Genotypic and survival characteristics of *Escherichia coli* phylogroup B2 from water

Angelin Gifta Henley Samuel

November 2018

A thesis submitted for the degree of Doctor of Philosophy of
The Australian National University

DECLARATION

I declare that this thesis titled "Genotypic and survival characteristics of *Escherichia coli* phylogroup B2 from water" is the outcome of my original research. To the best of my knowledge, this thesis contains no materials previously published or written by another person, except where due reference is made in the text. It contains no material which has been accepted for the award of any other degree or diploma in any university.

Angelin Gifta Henley Samuel

November 2018

ACKNOWLEDGMENTS

He gives beauty for ashes. Isaiah 61:3

I praise and thank my Lord and saviour Jesus Christ for his immense blessings upon me. Fixing my eyes on you Lord, I lose my fears and failures. I pray this thesis and my life glorifies God.

I owe my special thanks to my husband, Henley and my two sons, Hansel and Hadriel for their constant unconditional love, encouragement and motivation. Thank you, boys, for patiently understanding me at my hardest times and for the many many sacrifices you guys made just to make this journey of mine easy. This thesis wouldn't be possible without your support buddies. For this and much more, a big thank you.

I am truly grateful and indebt to my supervisor, Professor David Gordon, for guiding me in every step of the way on this Ph.D. journey. Thank you for patiently mentoring me from the moment I landed Australia and teaching me your wealth of research knowledge. Once again, I sincerely thank you Professor from the bottom of my heart.

My earnest thanks to Dr. Terry Neeman for her valuable advice on statistics and R programming. I would also like to thank Dr. Thy Truong for her technical support with mass spectrometry studies. Thanks also to the collaborators of this project, WaterNSW (Sydney Water), Sydney and Seqwater, Queensland.

Thank you to each of the members of the Gordon lab group past and present. Amar, Tijana, Julia and Sam thank you for teaching me many lab skills. Heli, Belinda, Sam, Buddhie, Charmalie, Mah, Judith, Upul and Truc, you have made working in the lab enjoyable and have become dear friends. Thank you Buddhie for the tea time chats, laughs and great friendship you shared. Thank you Charmalie for inspiring me with your hard work, faith and prayers.

I would like to thank my parents, dad (Robinson), mom (Johncy), mama (Muthiaraj) and athai (Ranjini) for all their love and support. Thank you for staying with us and helping us in times of need and building our family in Christ's love. My sister, Jeni and my siblings by heart Priestley anna, Deepa akka and Andrew, thank you for your unceasing support, pep talks and baby sittings. Hanniel, Ronel and Jordon, thank you for your love and witty jokes whenever I needed a good laugh. I also thank our house prayer group

members, Pastor Ravi, Ranji Aunty, Chitra Aunty, Mummy Aunty, Punitham Aunty, Debbie and Lisa for your prayers and yummy food deliveries.

This degree of Doctor of Philosophy is supported by an Australian Government Research Training Program (RTP) Scholarship. I gratefully acknowledge the financial support provided by the Australian Government.

ABSTRACT

Water is the most essential substance for life on earth. Hence, strict drinking water guidelines are framed to ensure the safety of drinking water supplies. For over a century, *Escherichia coli* has been used as the primary indicator of recent faecal contamination in water. *E. coli* is used as a faecal indicator bacteria (FIB) due to its high prevalence in the gut and faeces of humans, its ease of detection, and the assumption that *E. coli* cells quickly die once they leave the host. Recent population genetic studies are challenging these assumptions and suggest that *E. coli* is a versatile species and that some strains have adapted to the external environment or may even have become free-living without any association with the human host. As such, *E. coli* as FIB is increasingly questioned. Additionally, water industry has been trying to find methods to identify the source of faecal inputs to waterways, including typing of *E. coli* that have been isolated from the water.

For this purpose, first, the prevalence of human associated *E. coli* strains in water samples from various catchments across Sydney and southeast Queensland regions was investigated. Genotypic characterisation of this study revealed that the four predominantly human associated Sequence Types (ST)s (73, 95, 131, and ST69) represent less than 1% of the total *E. coli* isolates evaluated. This indicates that the *E. coli* in these drinking water sources are either non-human in origin or not recently contaminated with human activities. Second, a comparative genomics approach was used to contrast host and environmental isolates of *E. coli* to determine the extent to which the variable gene content of isolates from these two environments differed. This study showed two distinct clusters, one predominantly human associated and another native vertebrate animal associated. The environmental water isolates were equally distributed between the two clusters. The results hence suggest that not all *E. coli* from environment are human associated but may originate from animals as well. Third, an experiment was conducted to compare the survival pattern of both host and environmental isolates of *E. coli* in different water treatment types such as heat sterilisation and filter sterilisation and investigated on the variable gene content of these isolates to better understand the variation in survival with respect to each treatment. This study results suggested that contrary to the expectations that *E. coli* has poor survival in water, some went dormant achieving viable but non-culturable state (VBNC), exclusively in heat sterilised water, and some *E. coli* strains survived for extended periods in both water treatments. Further

evaluation showed that the among strain variation observed has an underlying genetic component.

Hence, to best consider *E. coli* as a FIB, all the investigations indicate that the difference within *E. coli* need to be considered and further characterised to differentiate true human *E. coli* and *E. coli* from other non-human sources. Overall, the results of these studies contribute towards understanding the limitations of using *E. coli* as an indicator of recent faecal pollution in water.

TABLE OF CONTENTS

CHAPTER 1

General Introduction

Escherichia coli- Definition

Escherichia coli (*E. coli*) is a versatile species that belongs to the family *Enterobacteriaceae*. It is named after a German pediatrician Theodor Escherich who, in 1884, was studying the microbial community of the gastrointestinal (GI) tract of human neonates and infants (Shulman et al., 2007). In the laboratory, isolates that ferments lactose and lysine decarboxylase positive, citrate negative, and produce indole are identified as *E. coli*. It is a rod shaped, coliform bacterium, and facultative anaerobe, commonly found in the lower intestine of humans and other warm-blooded animals as a commensal microorganism. *E. coli* is one of the first bacterial species to initially colonise the intestine of newborns, rapidly reaching a density of 10^9 CFU per gram of faeces and stabilises after 2 years at around 10^7 CFU per gram of faeces (Mitsuoka and Hayakawa, 1973). Although mostly harmless, some strains can cause serious extra intestinal infections and other food and water related diseases (Hartl and Dykhuizen, 1984). The first pathogenic strain was identified in the early 1940s in association with an outbreak of infantile diarrhea (Bray, 1945).

E. coli Genomics

Since this bacterium have simple molecular genetic systems and can be grown readily in the laboratory, *E. coli* remains as one of the best-studied prokaryotic model organisms. Recent genome sequencing projects have shown that a typical *E. coli* genome consists of about 4700 genes with about 2000 of these genes being common to all *E. coli* strains (core genome). Collectively, the species has a total gene pool (pan genome) of > 90 000 genes illustrating a high level of plasticity in its genome and the significant impact of horizontal gene transfers (Touchon et al., 2009; Van Elsas et al., 2011; Land et al., 2015). Most recent estimates indicate that recombination is responsible for more base changes than mutation in *E. coli*. Theoretically, high recombination rate disrupts the clonal framework of a species. In *E. coli*, as recombination involves short fragments, they do not alter the global topology of the organism providing a clonal population structure (an array of stable lineages) and a clear phylogenetic signal (Tenaillon et al., 2010).

E. coli Substructure

E. coli has an extensive genetic substructure that are better defined in the recent years. Initially, based on the genetic substructure demonstrated by multi-locus enzyme electrophoresis (MLEE) studies, most isolates of *E. coli* can be assigned to one of the four main phylogenetic groups (A, B1, B2, D) (Whittam et al., 1983; Selander et al., 1987; Herzer et al., 1990). Phylogroup A and B1 are considered as sister groups, while phylogroup B2 is monophyletic and considered to represent the ancestral lineage of *E. coli* (Lecointre et al., 1998), whereas phylogroup D has at least two distinct clades (Gordon et al., 2008; Jaureguy et al., 2008). More recently, multi-locus sequence typing studies (MLST) have more precisely defined the subgroup structure of *E. coli*, and a number of new phylogroups have been identified and better defined. One such subgroup is known as phylogroup C and represents strains that are closely related to phylogroup B1 strains. Phylogroup E, of which the diarrheal pathogen O157:H7 is the best-known member, is a diverse group of strains more closely related to A, B1 and C strains. Phylogroup F strains are most closely related to strains belonging to phylogroup B2 (Tenaillon et al., 2010).

During the last decade, five novel cryptic clades (I to V) have also been identified in the lineage *Escherichia* (Walk et al., 2009) that are phenotypically indistinguishable from *E. coli* isolates *sensu stricto*. Most of the cryptic *Escherichia* clades are phylogenetically sufficiently distinct from one another and from *E. coli* that they should be considered as distinct species. However, detailed genomic analysis suggests that cryptic clade I strains should be considered as a phylogroup of *E. coli* (Walk et al.,2009; Luo et.al., 2011).

Many studies have shown that the distribution of strains belonging to these phylogroups is actually non-random. Among Australian vertebrates, phylogroup A may be isolated from any class of vertebrates. Phylogroup B1 is more abundant in ectothermic vertebrates, birds, and carnivorous mammals. In mammals with hindgut fermentation, the phylogroup B2 is more frequently detected than the rest of the phylogroups. Phylogroup D is rarely detected in ectothermic vertebrates but is more likely to be isolated from endothermic vertebrates (Gordon and Cowling, 2003). Phylogroup A and B1 strains are frequently observed in water, soil and sediment samples, while phylogroup B2 and D

strains are less likely to isolated from these environments (Picard et al., 1999; Walk et al., 2007; Touchon et al., 2009).

Strains of various phylogroup also vary in their phenotypic properties (Gordon, 2004) and their ability to cause disease (Johnson et al., 2001). Strains belonging to phylogroup B2 and to a lesser extent phylogroup D are often the cause for extra-intestinal infections in the parts of the bodies that are outside of the intestine such as neonatal meningitis and urinary tract infection. This could be explained by the high frequency of virulence traits that these strains harbor in their genome (Diard et al., 2010). Thus, strains belonging to the different phylogroups differ in their propensity to cause disease, ecological niche and life history characteristics (Bergthorssonm and Ochman, 1998; Johnson et al., 2001; Gordon and Cowling, 2003).

Transition from host to environment

Escherichia coli is one of the most versatile and hardy bacterial species. It alternates between its primary habitat, the gut of vertebrates, where it lives mostly as a commensal (Tenaillon et al., 2010), and its secondary habitat: water, soil, and sediment (Savageau, 1983). Each of these environments is equally complex differing markedly in their physical condition and nutrient availability. But little is known about *E. coli*'s ability to transition between these habitats. What is known is that *E. coli*'s growth rate and survival ability differ between two habitats. In its primary habitat, the temperature is usually stabilised around 37°C and normally about 10^6 cells of *E. coli* per gram of colon content is seen (Geldreich, 1976). While in its secondary habitat typical temperatures may be from 0°C-24°C and the net growth rate is usually negative for *E. coli*, with a net half-life of 1 day in water (Faust et al., 1975) and about 1-5 days in soil (Van Donsel et al., 1967).

Thus far two different hypotheses have been proposed that explains *E. coli*'s transition mechanism from primary to secondary habitat. First, a study by Savageau in 1983 suggested that the growth rate of *E. coli* in primary and secondary habitats are subject to selection of genes when the organism cycles between the two habitats. The study predicts that a typical *E. coli* spends half its life span inside its host and half its life span outside its host, and has dual molecular control mechanisms with two sets of regulatory systems, one that is turned on in the host and the other when the cell enters the external environment depending on demand. Second, Whittam in 1989 suggested that not all

strains respond equally well to transition and selection plays a dominant role in determining which cells survive the transition between the two hosts. For example, Phylogroups A, B1 and the Clade strains are better survivors in secondary habitats such as fresh water. Clade strains are rarely recovered from human faeces. In contrast, phylogroups B2 and D are dominant in primary habitat such as human gut (Power et al., 2005, Walk et al., 2007; Ratajczak et al., 2010; Clermont et al., 2011).

While *E. coli* in its primary habitat is generally termed as commensal or pathogenic, *E. coli* isolated from its secondary habitat can be classified in three ways: (i) Faecal *E. coli*- deposited only from faeces of humans and animals, (ii) Naturalised *E. coli*- Persistent faecal strains that are stress tolerant and adapted to environmental conditions due to mutations resulting in niche-specific adaptation (Chiang et al., 2011, Walk et al., 2007), and (iii) Free living *E. coli*- strains that are found mainly in water and other secondary habitat, whose persistence is independent of faecal inputs (Power et al., 2005).

E. coli stress response

It is conclusively known that *E. coli* primarily inhabit the intestinal tract of human and other warm-blooded animals where the conditions are highly favourable for its survival with abundant carbon/energy sources, moderate pH and temperature. In contrast, *E. coli* strains are also detected in the external environment where they are exposed to various challenging conditions such as low nutrients, fluctuating temperature, low pH, and high osmolarity (van Elsas et al., 2011). Although *E. coli* is detected in both the primary, intestinal habitat and the secondary, external environmental habitat such as soil and water, its life cycle and transition between host mechanisms, including its ability to withstanding stress are poorly understood. Various studies suggest that *E. coli* strains have many survival strategies to establish and adapt itself to the external environment. One such adaptation method is, when confronted with different stress conditions, *E. coli* transit from exponential growth phase to stationary phase (De Biase et al., 1999; Chubukov and Sauer, 2014; Pletnev et al., 2015). It is well established from a variety of studies that the sigma factor (σ^s) is the central regulator of general stress response in *E. coli* when exposed to harsh conditions and it is strongly expressed during stationary phase (Lange and Aronis, 1991; Patten et al., 2004). The (σ^s) factor also controls the expression of many other genes that are involved in combating response to various stress factors encountered

(Adnan et al., 2017). In 2015, Vital and colleagues described that environmental *E. coli* strains showed an increased expression of genes that are essential for stress resistance, through activation of sigma factor (σ^s). The typical phenotypic characteristics of (σ^s) dependant strain involving long-term survival of *E. coli* in a stressful environment are an adhesive extracellular matrix consisting of cellulose, curli fimbriae and other polysaccharides which make the colonies phenotype 'red-dry-rough (rdar)' on solid media containing Congo red dye (White et al., 2006; White et al., 2011; Romling, 2005). This phenotype is more frequently observed in B1 and cryptic clade strains than the strains belonging to any other phylogroup (Di Sante et al., 2016). Also, the genes associated with environmental adaptation includes genes involved in diol utilisation and lysozyme production, whereas the genes enhanced in enteric isolates includes genes associated with the transport and use of nutrients thought to be abundant in the gut, such as gluconate and fucose.

Another key survival strategy of *E. coli* is that some stressed population enters the Viable but Non-Culturable (VBNC) state (Xu et al., 1982). This state was first recognized in 1982 by Rita Colwell and her colleagues for *E. coli* and *Vibrio cholerae* in the aquatic environment (Oliver, 2016). Under this state, cells are not culturable (cannot form colonies) in laboratory conditions, but remain metabolically active. This evasion of laboratory detection poses a significant health risk as it provides a limitation on the use of *E. coli* as the faecal indicator organism when cells are in VBNC state (Ramamurthy et al., 2014). Viable but non-culturable cells differ from normal cells in their composition of cell walls and membrane integrity, adhesion properties, cellular morphology, metabolism, gene expression, physical and chemical resistances and virulence potential (Li et al., 2014). When compared with dead cells, which are typically characterized by a damaged membrane and being metabolically inactive, VBNC cells have an intact membrane and retain metabolic activity. Nonpathogenic and pathogenic strains of *E. coli* have demonstrated survival of sub-lethal environmental stress such as pH, salinity and adverse temperature conditions by entering into this unique VBNC state (Arana et al., 2004; Liu et al., 2009). Cells that enter the VBNC state can also exit this state and become fully culturable under certain change in conditions or when favourable growth condition returns (Oliver, 2005; Lin et. al., 2016). Generally, in vitro, the favourable growth condition is induced by addition of specific compounds that promote growth such as nutrient rich media or optimal growth promoting factors (Liu et al., 2009; Pinto et al.,

2011) or by removal of stress that causes the VBNC state (Ohtomo and Saito, 2001). Several other studies have shown that VBNC *E. coli* cells in food and water reverted to grow under favourable conditions and cause infections in human body (Dinu and Bach, 2011; Fakruddin et al., 2013; Ramamurthy et al., 2014; Zhao et al., 2017). This shift in the survival process is called 'resuscitation' (Oliver, 2016). Upon resuscitation, under favourable conditions, their return to an infectious state with the retention of virulence factors is considered to be a major public health concern (Li et al., 2014).

E. coli typing

It is a widely acknowledged fact that not all strains of *E. coli* are identical, they instead share a common set of characteristics and are related by descent. Despite a relatively high level of recombination, various approaches used to study the population structure of *E. coli* reinforce a strong clonal concept, allowing them to be delineated into phylogroups. Each clone/phylogroup varies in their ecological niche, lifestyle and propensity to cause disease (Gordon & Cowling, 2003). Consequently, strain typing and identification of *E. coli* phylogeny is of growing importance as it is a commensal and a major pathogen (Croxen & Finley, 2010), as well as a water quality indicator (van Elsas et al., 2011). In the pre-sequencing era, the clonal structure of *E. coli* was first supported by serotyping analysis using the combination of 173 O (Somatic) antigens, 80 K (Capsule) antigens, 56 H (flagellar) antigens (Tenaillon et al., 2010). Subsequent Multi Locus Enzyme Electrophoresis (MLEE) analysis showed that clones from temporally or geographically distinct hosts were identical. With the arrival of sequencing techniques, Multi Locus Sequence Typing (MLST) was employed for characterising the population genetics of each *E. coli* strains to clonal groups (Maiden et al., 1998). Currently, three MLST schemes are available hosted by Michigan State University, USA (Reid et al., 2000), Warwick Medical School, UK (Wirth et al., 2006), and the Pasteur Institute, France (Jaureguy et al., 2008). They use the unique combinations of several housekeeping genes nucleotide sequence to determine the Sequence Type (ST) of each isolate (Clermont et al., 2015). *E. coli* has a very extensive diversity in its Sequence Types (STs) (Gordon, 2010). Besides MLST, a two-locus approach called CH typing using the fumC and fimH genes has been developed with greater discriminatory power than MLST (Weissman et al., 2012). All these approaches, though good in phylogenetic discrimination, have their own drawbacks with respect to the costs associated with materials and labour. Neither

MLST nor CH typing on its own does not provide the important information about the phylogroups of *E. coli*.

The phylogroups of *E. coli* isolates are frequently assigned using a PCR based method known as the Clermont/ Triplex PCR, which was originally based on the presence or absence of *chuA* and *yjaA* genes and one DNA fragment (TspE4.C2). This method assigned *E. coli* strains to one of the four main phylogroups such as A, B1, B2 and D. With the expansion of knowledge based on the MLST database, the triplex PCR was updated to quadruplex PCR method with the addition of *arpA* genes, to allow delineation of seven phylogroups (A, B1, B2, C, D, E, F) (Clermont et al., 2013).

With recent advancements in high throughput sequencing technologies, Whole Genome Sequencing (WGS) is an emerging technique employed to investigate the genome structure of *E. coli* and its population ecology and epidemiology (Gilchrist et al., 2015). Although WGS has the potential to provide high discrimination and resolution for subtyping *E. coli* to individual investigators needs, a major drawback is that the investigator needs precise knowledge and understanding of the genomic content of strains for proper interpretation of the data acquired. Also, in the case of an occurrence of a cross-border outbreak, considerable collaboration between the clinical and reference laboratories is crucial to support the clinical management and control the spread of disease. For now, it is still not a fully established method for routine usage because of the cost, difficulty of performance and standardization, data processing and storage, and limitations on extractions of data relevant to individual researchers (Holmes et al., 2015). To avoid these drawbacks on time and cost by short and whole genome sequencing, many alternative PCR based methods that are rapid and universally adaptable for targeting and identifying clinically important lineages of *E. coli* based on clonal complexes have been developed (Clermont et al., 2015). Some PCR methods target single clones of clinical importance such as: *icd* and *gyrB* allele-specific PCR for ST 648 belonging to phylogroup F (Johnson et al., 2017), *fumC* allele-specific PCR for STc 69 belonging to phylogroup D (Johnson et al., 2004), detection of the *svg* marker for STc 95 of phylogroup B2 (Bidet et al., 2007), *mdh* and *gyrB* allele-specific PCRs for ST 131 of phylogroup B2 (Johnson et al., 2009), *pabB* allele-specific PCR for the ST 131-O25b clone of phylogroup B2 (Clermont et al., 2009), *trpA* allele-specific PCR for the ST 131-O16 clone of phylogroup B2 (Johnson et al., 2014). Recently, simpler and rapid detection of several clonal

complexes by using multiplex PCRs have been developed. They are, an allele-specific PCR for detection of the nine main dominant human-associated B2 STc (12, 14, 73, 95, 127, 131, 141, 144, 372) (Clermont et al., 2014) and region-specific PCR for four ST complexes 69, 73, 95, 131 known as 'Doumith PCR' (Doumith et al., 2015). These PCR methods are more favoured than MLST and WGS and are of growing significance as they identify the lineages that are dominantly human host associated and of clinical relevance, rapidly at a comparatively lower cost and manual effort.

E. coli in the human host

The bacterium *E. coli* is a normal inhabitant in the intestine of human and other mammalian hosts. It is predicted that humans are exposed to 10^4 cells of *E. coli* per gram of food ingested (Hartl and Dykhuizen, 1984). This high level of exposure would suggest for a very rapid turnover rate of *E. coli* strains in the gut. Despite the expectation, a few studies indicate that the clonal composition of *E. coli* is highly stable with one or two dominant strains that represent the majority (>90%) of the *E. coli* population in the gut, and the remaining 10% represent a large number of strains present at very low frequencies (Caugant et al., 1981, 1984; Alm et al., 2011; Gordon, 1997; Gordon and Lee, 1999).

Several studies indicate that the clonal composition of *E. coli* in a human host varies with host age, sex, and diet (Gordon et al., 2005; Vollmerhausen et al., 2011). Some strains persist and can be detected at regular intervals in a host for 1 to 2 years and in some cases, up to 5 years (Clermont et al., 2008; Reeves et al., 2011; Johnson et al., 2016). These strains are known as resident/ persistent strains. Other strains can be observed once or a few times at irregular intervals. These strains are called transient strains (Caugant et al., 1984; Gordon, 2001; Alm et al., 2011).

Various survey and experimental studies also indicate that the phylogroups of *E. coli* exhibit some degree of host preference. The relative abundance of *E. coli* isolates with respect to each phylogroup also differs across humans living in different parts of the world. Phylogroup A strains are more likely to be isolated from people living in tropical regions (Escobar-Paramo et al., 2004) while, phylogroup B2 and D are more predominant in populations with western diet (Tenaillon et al., 2010; Duriez et al., 2001; Escobar-Paramo et al., 2004; Smati et al., 2015). Also, B2 strains are frequently responsible for extra intestinal infections globally, having more virulence-associated genes (Leclerc et

al., 2001). These genes also contribute to the success of colonization and persistence for a long period of time within the human gut (Diard et al., 2010).

A few studies suggest that within host competition among phylogroups plays an important role in determining the diversity of strains found in the gut. When phylogroup B2 strains are numerically dominant, they tend to persist longer and fewer strains would be detected within the host than in hosts with other phylogroups dominance such as A, B1 or D (Moreno et al., 2009; Dixit et al., 2018). Furthermore, when phylogroup B2 strains are dominant, they tend not to co-occur with other phylogroups suggesting that B2 strains are not just numerically but also competitively dominant (Smati et al., 2013; Blyton et al., 2014). Studies also indicate that although phylogroup B2 is highly diverse comprising of many STs, in humans, only a few lineages such as STs 12, 14, 73, 95, 127, 131, 141, 144, 372 of phylogroup B2 and ST69 of phylogroup D are most likely to represent the total number of strains from extra intestinal sites and faeces, globally. Also, STs 69, 73, 95, 131 are more predominant and over-represented compared to the total lineages in humans (Clermont et al., 2014; Doumith et al., 2015; Riley, 2014; Le Gall et al., 2007). A study from the Gordon lab also indicated that ST 69 and ST 95 are frequently isolated from faeces of asymptomatic humans living in Australia but not from native Australian mammals (Gordon, 2013).

E. coli as a water quality indicator

Water contamination by faeces is an important human health issue and primary concern for water authorities as worldwide eighty-eight percent of diarrheal illness is attributed to unsafe water supplies and inadequate sanitation of water (Leclerc et al., 2001; Yongsi, 2010). Faeces can contribute bacteria, viruses or protozoa that can adversely affect public health. Direct measurement of all pathogenic microorganism is difficult, expensive and time-consuming. Water safety managers want to be able to quickly and inexpensively detect faecal contamination of water to prevent risks to consumers of water or people who recreate in the water. Hence Faecal Indicator Bacteria (FIB) are used to measure the possible presence of pathogenic bacteria and its associated public health risk. For many years *E. coli* has been considered as one of the main indicators used for this purpose as it is very easy to identify using conventional laboratory techniques (EPA, 1986; WHO, 2008). There are three main reasons or assumptions for considering *E. coli* as FIB in

fresh and drinking waters. Firstly, *E. coli* is considered specific to faecal materials from human and other warm-blooded animals, typically in high densities attaining 10^5 to 10^7 colony forming units per gram of faeces (Hartl and Dykhuizen, 1984; Slanetz and Bartley, 1957). Secondly, as well as possibly possessing virulence factors that are thought to be a potential threat themselves, they are also an indicator of other fecally derived pathogens and their associated potential diseases due to the suspicion that along with *E. coli* from faeces other faecal pathogenic organisms could be present (Ahmed et al., 2015). Thirdly, *E. coli* is generally considered to have poor survival ability in the external environment with the average lifespan of about 1 day in water and 5 days in soil (Faust et al., 1975; Ingle et al., 2011).

In recent decades, many publications have suggested that these assumptions are not always true. Understanding the life cycle and survival strategies of *E. coli* plays a critical role in using the species correctly as a water quality indicator or tracing the source of faecal contamination. When there is a known source of faecal contamination, such as sewage runoff, *E. coli* abundance in water is, of course, a good predictor of gastrointestinal illness (Cabelli et al., 1982). But, the presence of *E. coli* in the environment may not always reflect faecal contamination, nor pose a risk of enteric illness. Various studies worldwide from tropical (Carrillo et al., 1985; Hardina and Fujioka 1991; Jimenez et al., 1989; Lopez-Torres et al., 1987), subtropical (Power et al., 2005; Solo-Gabriela et al., 2000), and temperate (Alm et al., 2006; Byappanahalli et al., 2007; Ksoll et al., 2007) regions detected high densities of *E. coli* in the external environment but concluded that those isolates were not associated with faecal source contamination. Hence, two observations based on these recent studies challenge *E. coli* to be considered as a FIB. The first, *E. coli* populations in secondary habitats are very large and diverse compared to the *E. coli* detected from human and warm-blooded vertebrates, suggesting that they are not faecal derived (Savageau, 1983). A study by Ishii and colleagues, 2006, on *E. coli* isolates from the soils of Minnesota, United States, reported that they found 'naturalized' *E. coli* isolates that were distinct from each other and from water and animal faecal isolates of the same region. Similarly, studies on a collection of *E. coli* isolates by Walk and his colleagues (Walk et al., 2007) from 6 beaches across Great Lake Watershed with known urban, industrial and agricultural runoff found a single unique genotype that was independently sampled 7 times from different water columns and sand cores across 5 beaches. They suggested that recovery

of this unique genotype could be of persistent *E. coli* isolates that are adapted to life outside the GI tract. Another key observation is that these strains can replicate and reach high densities that would be thought to represent recent faecal contamination under favourable condition outside mammalian hosts even in the absence of regular faecal input (Blyton and Gordon, 2017). A study on *E. coli* bloom strains (detected in excess of 10^4 CFU/100ml) by Power and colleagues between 2002 and 2004 on Lake Burragorang and Lake Burley Griffin, Australia, concluded that those *E. coli* bloom strains have evolved a free- living lifestyle and are highly unlikely to be originated from faecal contamination (Power et al., 2005). Hence, these observations suggest that *E. coli* therefore may not always be a good indicating measure of faecal pollution and the risk associated to faecal contamination.

Research Aims

Although *E. coli* is one of the well-studied model organisms, especially in the context of being a "normal inhabitant" of the GI tract of human and animals, very little is known about the fate/ survival characteristics of *E. coli* cells during the transition process from the GI tract of a primary host to the secondary habitat such as soil, water, sediment. Gaining an understanding by comparing and contrasting the phenotypic and genotypic characteristics of host associated *E. coli* strains isolated from the secondary environment is essential if we are to use *E. coli* as an indicator for faecal contamination in water. The highest risk to human health is attributed to human faecal contamination in water. Faeces from livestocks or domestic animals or native animals or birds is regarded as lesser health risk than from human faeces as it is likely to contain fewer pathogens that can infect humans. Since, generally the biggest concern of water authorities and the public are about human faecal contamination that is most likely to represent a human health risk, Phylogroup B2 and D strains of *E. coli* are the apt target for this concern and the focus of this thesis as (i) they represent more than 50% of the isolates from humans living in industrialised countries, (ii) they are thought to have poor survival ability in water but prolonged and persistent ability to survive in humans, (iii) although the phylogroup B2 have great diversity in its ST, only a few STs dominate and persist in humans and these dominant STs are uncommon in other vertebrates and livestocks, (iv) simple PCR based detection methods are available to detect the dominant STs rapidly, and (v) they possess a contrasting profile of virulence and commensalism in the human gut. Another

advantage in studying the environmental survival aspects of *E. coli* belonging to these phylogroups is that it will widen our understanding of its ecology, transition strategies between primary and secondary habitat, and its suitability for water quality monitoring.

The first chapter of this thesis surveyed the *E. coli* strains belonging to phylogroup B2 and D from various catchments and dams of Sydney and Queensland. The *E. coli* isolates with B2 and D phylogroup were sequence typed and genetically characterised for their host association. This enabled the differentiation between the host associated and environmentally adapted isolates within the phylogroup B2 and D at the clonal level. Jaureguy and colleagues in 2008 proposed that for clinically significant bacteria, studying them by clonal complexes and sequence types could possibly be more applicable than by their higher classification by phylogroups. Hence even among isolates with phylogroups B2 and D, the sequence types ST 73, ST 95, ST 131, ST 69 are focused on, as they are more dominant in humans especially from clinical samples and are of epidemiological importance. The second chapter of this thesis aimed to compare and contrast the phylogroup B2 isolates from water with existing isolates collected from humans and other native vertebrates at the genomic level. The third chapter examined the survival ability of host associated phylogroup B2 isolates in water. This understanding of its survival ability is critically important in the safe management of water supplies. Finally, I conclude the importance of my work in relation to *E. coli* as a water quality indicator as the results shed light on the ecology and life history characteristics of the organism.

References

Adnan, M. et al., 2017. Role of *bolA* and *rpoS* genes in biofilm formation and adherence pattern by *Escherichia coli* K-12 MG1655 on polypropylene, stainless steel, and silicone surfaces. *Acta Microbiologica et Immunologica Hungarica,* 64(2), pp. 179-189.

Alm, E. W., Burke, J. & Hagan, E., 2006. Persistence and potential replication of the fecal indicator bacteria, *Escherichia coli*, in the shoreline sand at Lake Huron. *Journal of Great Lakes Research ,* Volume 32, pp. 401-405.

Alm, E. W., Walk, S. T. & Gordon, D. M., 2011. The niche of *Escherichia coli*. In: S. T. Walk & P. C. H. Feng, eds. *Population Genetics of Bacteria.* Washington, DC, USA: ASM Press, pp. 107-123.

Arana, I. et al., 2004. Relationship between *E. coli* cells and the surrounding medium during survival processes. *Antonie van Leeuwenhoek,* 86(2), pp. 189-199.

Bergthorsson, U. & Ochman, H., 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Molecular Biology and Evolution,* 15(1), pp. 6-16.

Bidet, P. et al., 2007. Detection and Identification by PCR of a Highly Virulent Phylogenetic Subgroup among Extraintestinal Pathogenic *Escherichia coli* B2 Strains. *Applied and Environmental Microbiology,* 73(7), p. 2373–2377.

Blyton, M. D. J. et al., 2014. Not all types of host contacts are equal when it comes to *E. coli* transmission. *Ecology Letters ,* Volume 17, pp. 970-978.

Blyton, M. D. J. & Gordon, D. M., 2017. Genetic Attributes of *E. coli* Isolates from Chlorinated Drinking Water. *PLoS ONE,* 12(1), p. e0169445.

Bray, J., 1945. Isolation of antigenically homogeneous strains of *Bacterium coli neopolitanum* from summer diarrhoea of infants. *The Journal of Pathology and Bacteriology,* 57(2), pp. 239-247.

Byappanahalli, M. N. et al., 2007. Population structure of Cladophora-borne *Escherichia coli* in nearshore water of Lake Michigan. *Water Research,* 41(16), pp. 3649-3654.

Cabelli, V. J., Dufour, A. P., McCabe, L. J. & Levin , M. A., 1982. Swimming-associated gastroenteritis and water quality. *American Journal of Epidemiology,* 115(4), pp. 606-616.

Carrillo, M., Estrada, E. & Hazen, T. C., 1985. Survival and enumeration of the fecal indicators *Bifidobacterium adolescenti*s and *Escherichia coli* in a tropical rain forest watershed. *Applied and Environmental Microbiology,* 50(2), pp. 468-476.

Caugant, D. A., Levin, B. R. & Selander, R. K., 1981. Genetic diversity and temporal variation in the *E. coli* population of a human host. *Genetics,* 98(3), pp. 467-490.

Caugant, D. A., Levin, B. R. & Selander, R. K., 1984. Distribution of multilocus genotypes of *Escherichia coli* withim and between host families. *The Journal of Hygiene,* 92(3), pp. 377-384.

Chiang, S. M., Dong, T., Edge, T. A. & Schellhorn, H. E., 2011. Phenotypic Diversity Caused by Differential RpoS Activity among Environmental *Escherichia coli* Isolates. *Applied and Environmental Microbiology,* 77(22), p. 7915–7923.

Chubukov, V. & Sauer, U., 2014. Environmental dependence of stationary-phase metabolism in *Bacillus subtilis* and *Escherichia coli*. *Applied and Environmental Microbiology,* 80(9), pp. 2901-2909.

Clermont, O. et al., 2014. Development of an allele-specific PCR for *Escherichia coli* B2 sub-typing, a rapid and easy to perform substitute of multilocus sequence typing. *Journal of Microbiological Methods,* Volume 101, pp. 24-27.

Clermont, O., Christenson, J. K., Denamur, E. & Gordon, D. M., 2013. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports,* 5(1), pp. 58-65.

Clermont, O. et al., 2009. Rapid detection of the O25b-ST131 clone of *Escherichia coli* encompassing the CTX-M-15-producing strains. *The Journal of Antimicrobial Chemotheraphy,* 64(2), pp. 274-277.

Clermont, O., Gordon, D. & Denamur, E., 2015. Guide to the various phylogenetic classification schemes for *Escherichia coli* and the correspondence among schemes. *Microbiology,* 161(5), pp. 980-988.

Clermont, O. et al., 2011. Characterization of the cryptic Escherichia lineages: rapid identification and prevalence. *Environmental Microbiology,* 13(9), pp. 2468-2477.

Clermont, O. et al., 2008. Evidence for a human-specific *Escherichia coli* clone. *Environmental Microbiology,* 10(4), pp. 1000-1006.

Croxen, M. A. & Finlay, B. B., 2010. Molecular mechanisms of *Escherichia coli* pathogenicity. *Nature Reviews Microbiology,* 8(1), pp. 26-38.

De Biase, D., Tramonti, A., Bossa, F. & Visca, P., 1999. The response to stationary-phase stress conditions in *Escherichia coli*: role and regulation of the glutamic acid decarboxylase system. *Molecular microbiology,* 32(6), pp. 1198-1211.

Di Sante, L. et al., 2016. Multicellular behavior of environmental *Escherichia coli* isolates 1 grown under 2 nutrient-poor and low-temperature conditions. *Microbiological Research,* Volume 210, pp. 43-50.

Diard, M. et al., 2010. Pathogenicity-associated islands in extraintestinal pathogenic *Escherichia coli* are fitness elements involved in intestinal colonization. *Journal of Bacteriology,* 192(19), pp. 4885-4893.

Dinu, L. D. & Bach, S., 2011. Induction of viable but nonculturable *Escherichia coli* O157:H7 in the phyllosphere of lettuce: a food safety risk factor. *Applied and Environmental Microbiology,* 77(23), pp. 8295-8302.

Dixit , O. V. A., O'Brien, C. L., Pavli, P. & Gordon, D., 2018. Within-host evolution versus immigration as adeterminant of *Escherichia coli* diversity in the human gastrointestinal tract. *Environmental Microbiology* , 20(3), pp. 993-1001.

Doumith, M. et al., 2015. Rapid Identification of Major *Escherichia coli* Sequence Types Causing Urinary Tract and Bloodstream Infections. *Journal of Clinical Microbiology,* 53(1), pp. 160-166.

Duriez, P. et al., 2001. Commensal *Escherichia coli* are phylogenetically distributed among geographically distinct human populations. *Microbiology,* 147(6), pp. 1671-1676.

EPA (Environmental Protection Agency) Report of Task Force on Guide Standard and Protocol for Testing Microbiological Water Purifiers. (1986).

Escobar-Páramo, P. et al., 2004. A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. *Molecular Biology and Evolution,* 21(6), pp. 1085-1094.

Escobar-Páramo, P. et al., 2004. Large-scale population structure of human commensal *Escherichia coli* isolates. *Applied and Environmental Microbiology,* 70(9), pp. 5698-5700.

Fakruddin, M., Mannan, K. S. B. & Andrews, S., 2013. Viable but Nonculturable Bacteria: Food Safety and Public Health Perspective. *ISRN Microbiology,* Volume 2013, p. 703813.

Faust, M. A., Aotaky, A. E. & Hargadon, M. T., 1975. Effect of Physical Parameters on the In Situ Survival of *Escherichia coli* MC-6 in an Estuarine Environment. *Applied Microbiology,* 30(5), pp. 800-806.

Geldreich,, E. E., 1976. Fecal coliform and fecal streptococcus density relationships in waste discharges and receiving waters. 6(4), pp. 349-369.

Gilchrist, C. A. et al., 2015. Whole-Genome Sequencing in Outbreak Analysis. Clinical Microbiology Reviews, 28(3), pp. 541-563.

Gordon, D. M., 1997. The genetic structure of *Escherichia coli* populations in feral house mice. *Microbiology,* Volume 143, pp. 2039-2046.

Gordon, D. M., 2001. Geographical structure and host specificity in bacteria and the implications for tracing the source of coliform contamination. *Microbiology,* Volume 147, pp. 1079-1085.

Gordon, D. M., 2004. The influence of ecological factors on the distribution and genetic structure of *Escherichia coli*.. In: D. A. Rasko, ed. *Escherichia coli and Salmonella*

typhimurium: cellular and molecular biology. Washington, D. C.: American Society for Microbiology.

Gordon, D. M., 2010. Strain typing and the ecological structure of *Escherichia coli*. *Journal of AOAC International,* 93(3), pp. 974-984.

Gordon, D. M., 2013. The ecology of *Escherichia coli*. In: M. S. Donnenberg, ed. *Escherichia coli: Pathotypes and Principles of Pathogenesis.* Maryland, USA: Elsevier Inc., pp. 3-20.

Gordon, D. M., Clermont, O., Tolley, H. & Denamur, E., 2008. Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environmental Microbiology,* 10(10), p. 2484–2496.

Gordon, D. M. & Cowling, A., 2003. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology,* Volume 149, p. 3575–3586.

Gordon, D. M. & Lee, J., 1999. The genetic structure of enteric bacteria from Australian mammals. *Microbiology,* Volume 145, pp. 2673-2682.

Gordon, D. M., Stern, S. E. & Collignon, P. J., 2005. Influence of the age and sex of human hosts on the distribution of *Escherichia coli* ECOR groups and virulence traits. *Microbiology,* 151(1), pp. 15-23.

Hardina, C. M. & Fujioka, R. S., 1991. Soil: The environmental source of *Escherichia coli* and enterococci in Hawaii's streams. *Environmental Toxicology,* 6(2), pp. 185-195.

Hartl, D. L. & Dykhuizen, D. E., 1984. The population genetics of *Escherichia coli*. *Annual Review of Genetics* , Volume 18, pp. 31-68.

Herzer, P. J., Inouye, S., Inuye, M. & Whittam, T. S., 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. *Journal of Bacteriology,* 172(11), pp. 6175-6181.

Holmes, A. et al., 2015. Utility of Whole-Genome Sequencing of *Escherichia coli* O157 for Outbreak Detection and Epidemiological Surveillance. *Journal of Clinical Microbiology*, 53(11), pp. 3565-3573.

Ingle, D. J. et al., 2011. Biofilm Formation by and Thermal Niche and Virulence Characteristics of Escherichia spp. *Applied and Environmental Microbiology*, 77(8), pp. 2695-2700.

Ishii, S., Ksoll, W. B., Hicks, R. E. & Sadowsky, M. J., 2006. Presence and Growth of Naturalized *Escherichia coli* in Temperate Soils from Lake Superior Watersheds. *Applied and Environmental Microbiology*, 72(1), pp. 612-621.

Jaureguy , F. et al., 2008. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BMC Genomics*, Volume 9, p. 560.

Jiménez, L., Muñiz, I., Toranzos, G. A. & Hazen, T. C., 1989. Survival and activity of Salmonella typhimurium and *Escherichia coli* in tropical freshwater. *The Journal of Applied Bacteriology*, 67(1), pp. 61-69.

Johnson, J. R. et al., 2014. Rapid and specific detection, molecular epidemiology, and experimental virulence of the O16 subgroup within *Escherichia coli* sequence type 131. *Journal of Clinical Microbiology*, 52(5), p. 1358–1365.

Johnson, J. R., Delavari, P., Kuskowski, M. & Stell, A. L., 2001. Phylogenetic Distribution of Extraintestinal Virulence-Associated Traits in *Escherichia coli*. *The Journal of Infectious Diseases*, 183(1), pp. 78-88.

Johnson, J. R., Johnston, B. D. & Gordon, D. M., 2017. Rapid and Specific Detection of the *Escherichia coli* Sequence Type 648 Complex within Phylogroup F. *Journal of Clinical Microbiology*, 55(4), pp. 1116-1121.

Johnson, J. R. et al., 2016. *Escherichia coli* Sequence Type 131 H30 Is the Main Driver of Emerging Extended-Spectrum-β-Lactamase- Producing *E. coli* at a Tertiary Care Center. *mSphere*, 1(6), pp. e00314-16.

Johnson, J. R. et al., 2009. Epidemic clonal groups of *Escherichia coli* as a cause of antimicrobial-resistant urinary tract infections in Canada, 2002 to 2004. *Antimicrobial Agents and Chemotherapy*, 53(7), pp. 2733-2739.

Johnson, J. R., Owens, K., Manges, A. R. & Riley, L. W., 2004. Rapid and Specific Detection of *Escherichia coli* Clonal Group A by Gene-Specific PCR. *Journal of Clinical Microbiology,* 42(6), pp. 2618-2622.

Ksoll, W. B., Ishii, S., Sadowsky, M. J. & Hicks, R. E., 2007. Presence and Sources of Fecal Coliform Bacteria in Epilithic Periphyton Communities of Lake Superior. *Applied and Environmental Microbiology,* 73(12), pp. 3771-3778.

Land, M. et al., 2015. Insights from 20 years of bacterial genome sequencing. *Functional and Integrative Genomics,* 15(2), pp. 141-161.

Lange, R. & Hengge-Aronis, R., 1991. Identification of a central regulator of stationary-phase gene expression in *Escherichia coli*. *Molecular Microbiology,* 5(1), pp. 49-59.

Le Gall, T. et al., 2007. Extraintestinal Virulence Is a Coincidental By-Product of Commensalism in B2 Phylogenetic Group *Escherichia coli* Strains. *Molecular Biology and Evolution,* 24(11), pp. 2373-2384.

Leclerc, H., Mossel, D. A., Edberg, S. C. & Struijk, C. B., 2001. Advances in the bacteriology of the coliform group: their suitability as markers of microbial water safety. *Annual Review of Microbiology,* Volume 55, pp. 201-234.

Lecointre, G., Rachdi, L., Darlu, P. & Denamur, E., 1998. *Escherichia coli* molecular phylogeny using the incongruence length difference test. *Molecular Biology and Evolution,* 15(12), pp. 1685-1695.

Li, L. et al., 2014. The importance of the viable but non-culturable state in human bacterial pathogens. *Frontiers in Microbiology,* 5(258).

Lin , Y. W. et al., 2016. Bacterial regrowth in water reclamation and distribution systems revealed by viable bacterial detection assays. *Chemosphere,* Volume 144, pp. 2165-2174.

Liu, Y. et al., 2009. Induction of *Escherichia coli* O157:H7 into the viable but non-culturable state by chloraminated water and river water, and subsequent resuscitation. *Environmental Microbiology Reports,* 1(2), pp. 155-161.

Lopez-Torres, A. J., Hazen, T. C. & Toranzos, G. A., 1987. Distribution and in situ survival and activity of *Klebsiella pneumoniae* and *Escherichia coli* in a tropical rain forest watershed. *Current Microbiology,* 15(4), pp. 213-218.

Luo, C. et al., 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *PNAS,* 108(17), p. 7200–7205.

Maiden, M. C. J. et al., 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences,* 95(6), pp. 3140-3145.

Mitsuoka, T. & Hayakawa, K., 1973. The fecal flora in man. I. Composition of the fecal flora of various age groups. *Zentralblatt fur Bakteriologie, Parasitenkunde, Infektionskrankheiten und Hygiene,* 223(2), pp. 333-342 (In German).

Moreno, E. et al., 2009. Structure and urovirulence characteristics of the fecal *Escherichia coli* population among healthy women. *Microbes and Infection,* 11(2), pp. 274-280.

Ohtomo, R. & Saito, M., 2001. Increase in the Culturable Cell Number of *Escherichia coli* during Recovery from Saline Stress: Possible Implication for Resuscitation from the VBNC State. *Microbial Ecology,* 42(2), pp. 208-214.

Oliver, J. D., 2005. The Viable but Nonculturable State in Bacteria. *The Journal of Microbiology,* 43(5), pp. 93-100.

Oliver, J. D., 2016. The Viable but Nonculturable State for Bacteria: Status Update. *Microbe,* 11(4), pp. 159-164.

Patten, C. L. et al., 2004. Microarray analysis of RpoS-mediated gene expression in *Escherichia coli* K-12. *Molecular Genetics and Genomics,* 272(5), pp. 580-591.

Picard, B. et al., 1999. The Link between Phylogeny and Virulence in *Escherichia coli* Extraintestinal Infection. *Infection and Immunity,* 67(2), pp. 546-553.

Pinto, D., Almeida, V., Almeida Santos, M. & Chambel, L., 2011. Resuscitation of *Escherichia coli* VBNC cells depends on a variety of environmental or chemical stimuli. *Journal of Applied Microbiology,* 110(6), pp. 1601-1611.

Pletnev, P. et al., 2015. Survival guide: *Escherichia coli* in the stationary phase. *Acta Naturae,* 7(4), pp. 22-33.

Power, M. L. et al., 2005. Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. *Environmental Microbiology,* 7(5), pp. 631-640.

Ramamurthy, T., Ghosh, A., Pazhani, G. P. & Shinoda, S., 2014. Current perspectives on viable but non-culturable (VBNC) pathogenic bacteria. *Frontiers in Public Health,* Volume 2.

Ratajczak, M. et al., 2010. Influence of hydrological conditions on the *Escherichia coli* population structure in the water of a creek on a rural watershed. *BMC Microbiology,* Volume 10, p. 222.

Reeves, P. R. et al., 2011. Rates of Mutation and Host Transmission for an *Escherichia coli* Clone over 3 Years. *PLoS ONE,* 6(10), p. e26907.

Reid, S. D. et al., 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature,* 406(6791), pp. 64-67.

Riley, L. W., 2014. Pandemic lineages of extraintestinal pathogenic *Escherichia coli*. *Clinical Microbiology and Infection,* Volume 20, pp. 380-390.

Romling, U., 2005. Characterization of the rdar morphotype, a multicellular behaviour in Enterobacteriaceae. *Cellular and Molecular Life Sciences,* 62(11), pp. 1234-1246.

Savageau, 1983. *Escherichia coli* habitats, cell types, and molecular mechanisms of gene control. *The American Naturalist,* 122(6), pp. 732-744.

Selander, R. K. et al., 1987. Population genetics of pathogenic bacteria. *Microbial Pathogenesis*, 3(1), pp. 1-7.

Shulman, S. T., Friedmann, H. C. & Sims, R. H., 2007. Theodor Escherich: the first pediatric infectious diseases physician?. *Clinical Infectious Diseases* , 45(8), p. 1025–1029.

Slanetz, L. W. & Bartley, C. H., 1957. Numbers of Enterococci in water, sewage, and feces determined by the membrane filter technique with an improved medium. *Journal of Bacteriology*, 74(5), pp. 591-595.

Smati, M. et al., 2015. Quantitative analysis of commensal *Escherichia coli* populations reveals host-specific enterotypes at the intra-species level. *Microbiology*, 4(4), pp. 604-615.

Smati, M. et al., 2013. Real-Time PCR for Quantitative Analysis of Human Commensal *Escherichia coli* Populations Reveals a High Frequency of Subdominant Phylogroups. *American Society for Microbiology*, 79(16), pp. 5005-5012.

Solo-Gabriele, H. M., Wolfert, M. A., Desmarais, T. R. & Palmer, C. J., 2000. Sources of *Escherichia coli* in a coastal subtropical environment. *Applied and Environmental Microbiology*, 66(1), pp. 230-237.

Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E., 2010. The population genetics of commensal *Escherichia coli*. *Natures Reviews Microbiology*, Volume 8, pp. 207-217.

Touchon, M. et al., 2009. Organised genome dynamics in the *Escherichia coli* species results in high diverse adaptive paths. *PLOS Genetics*, 5(1), p. 5:e1000344.

Van Donsel, D. J., Geldriech, E. E. & Clarke, N. A., 1967. Seasonal Variations in Survival of Indicator Bacteria in Soil and Their Contribution to Storm-water Pollution. *Applied Microbiology*, 15(6), pp. 1362-1370.

van Elsas, J. D., Semenov, A. V., Costa, R. & Trevors, J. T., 2011. Survival of *Escherichia coli* in the environment: fundamental and public health aspects. *The ISME Journal*, Volume 5, pp. 173-183.

Vital, M. et al., 2015. Gene expression analysis of *E. coli* strains provides insights into the role of gene regulation in diversification. *The ISME Journal,* 9(5), pp. 1130-1140.

Vollmerhausen, T. L. et al., 2011. Population structure and uropathogenic virulence-associated genes of faecal *Escherichia coli* from healthy young and elderly adults. *Journal of Medical Microbiology,* 60(5), pp. 574-581.

Walk, S. T. et al., 2007. Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. *Environmental Microbiology,* 9(9), pp. 2274-2288.

Walk, S. T. et al., 2009. Cryptic lineages of the genus *Escherichia*. *Applied and Environmental Microbiology,* 75(20), p. 6534–6544.

Ahmed, W. et al., 2015. Assessment of Genetic Markers for Tracking the Sources of Human Wastewater Associated *Escherichia coli* in Environmental Waters. *Environmental Science and Technology*, 49(15), pp. 9341-9346.

Weissman, S. J. et al., 2012. High-resolution two-locus clonal typing of extraintestinal pathogenic *Escherichia coli*. *Applied and Environmental Microbiology,* 78(5), pp. 1353-1360.

White, A. P. et al., 2006. Thin aggregative fimbriae and cellulose enhance long-term survival and persistence of Salmonella. *Journal of Bacteriology,* 188(9), pp. 3219-3227.

White, A. P. et al., 2011. Intergenic Sequence Comparison of *Escherichia coli* Isolates Reveals Lifestyle Adaptations but Not Host Specificity. *Applied and Environmental Microbiology,* 77(21), pp. 7620-7632.

Whittam, T. S., 1989. Clonal dynamics of *Escherichia coli* in its natural habitat. *Antonie Van Leeuwenhoek,* 55(1), pp. 23-32.

Whittam, T. S., Ochman, H. & Selander, R. K., 1983. Multilocus genetic structure in natural populations of *Escherichia coli*. *PNAS,* 80(6), pp. 1751-1755.

WHO (World Health Organization). Guidelines for Drinking-water Quality, Incorporating 1st and 2nd Addenda, Volume 1, Recommendations, 3rd ed.; WHO: Geneva, Switzerland (2008).

Wirth, T. et al., 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Molecular Microbiology,* 60(5), pp. 1136-1151.

Xu, H. S. et al., 1982. Survival and Viability of Nonculturable *Escherichia coli* and Vibrio cholerae in the Estuarine and Marine Environment. *Microbial Ecology,* Volume 8, pp. 313-323.

Yongsi, H. B. N., 2010. Suffering for Water, Suffering from Water: Access to Drinking-water and Associated Health Risks in Cameroon. *Journal of Health, Population and Nutrition,* 28(5), pp. 424-435.

Zhao, X. et al., 2017. Current Perspectives on Viable but Non-culturable State in Foodborne Pathogens. *Frontiers in Microbiology,* 8(580).

CHAPTER 2

Frequency of human associated *E. coli* lineages in water samples from two regions of Australia

Introduction

Water is an elixir of our life. The Australian Drinking Water Guidelines (the ADWG) provides strict procedures for the management of good quality drinking water supplies. Monitoring the microbial quality of water by testing for *Escherichia coli* as a faecal indicator organism is one of the important quality assurance tests framed in ADWG to ensure the safety of public health (ADWG, 2011). There are two main reasons for considering *E. coli* as faecal indicator organism, firstly, *E. coli* is regarded to be present in large numbers and specific to the faeces of humans and other warm-blooded animals, and secondly, they are considered to be a transient member of the microbial community of water due to their poor survival (WHO, 1993; Edberg et al., 2000; NHMRC-ARMCANZ, 1996). *E. coli* is a well-known commensal of the mammalian gut (Tenallion et al., 2010). However, it can occasionally be responsible for various intestinal and extra intestinal infections as well (Johnson, 2002; Cabral, 2010). *E. coli* enters water bodies through various sources: animal waste from native and farm animals, sewage overflows, polluted stormwater runoff, agricultural runoff (manure, fertiliser) and industrial wastes. Although *E. coli* has multiple potential sources of origin for contamination, faecal contamination from humans represent the greatest risk of waterborne pathogenic diseases (Regli et al., 1991; Harwood et al., 2014).

E. coli as a species is very versatile and exhibits a strong clonal genetic structure (Selander and Levin, 1980; Desjardins et al., 1995), with the majority of strains belonging to one of the four main phylogenetic groups - A, B1, B2, D (Herzer et al., 1990; Wirth et al., 2006). Each phylogenetic group differs in their phenotypic and genotypic traits and has different ecological niches and life history characteristics (Gordon and Cowling, 2008; Alm et al., 2011). Most of the *E. coli* strains isolated from extra intestinal body sites are phylogroup B2 and to a lesser extent phylogroup D, and these strains frequently encode a diversity of virulence traits (Bingen et al., 1998; Johnson et al., 2001; Johnson, 2002). Interestingly, the majority of faecal samples from asymptomatic humans living in developed countries

harbor phylogroup B2 strains as the predominant commensals, although their frequency varies with diet, host age and sex (Gordon et al., 2005; Nowrouzian et al., 2006; Escobar-Paramo et al., 2006; Le Gall et al., 2007; Smati et al., 2015; Gordon et al., 2015).

Although phylogroup B2 is genetically highly diverse with hundreds of subgroups or STs, MLST studies have revealed that a relatively small number of lineages are isolated from humans (ST 12, ST 14, ST 73, ST 95, ST 127, ST 131, ST 141,ST 144, ST 372) (Massot et al., 2016; Doumith et al., 2015; Mahjoub-Messai et al., 2011; Gibreel et al., 2012; Bert et al., 2010; Kallonen et al., 2017). Of these, three STs (ST 73, ST 95, ST 131) represent the majority of B2 strains detected in human faeces and extra intestinal sites. Together, the evidence suggests that these lineages exhibit considerable host specificity (Le Gall et al., 2007; Clermont et al., 2011; Day et al., 2016; Gordon et al., 2017). Similarly, a clinically significant lineage belonging to phylogroup D is clonal group A (CGA) corresponding to ST 69 (Bidet et al., 2007). Locally, in healthy asymptomatic humans living in Australia, Clonal groups ST 69 and ST 95 are frequently isolated from faeces, yet very rarely detected in native Australian mammals (Gordon, 2013; Gordon et al., 2017).

Given that human faecal contamination of water bodies likely represents the greatest risk to human health, then these largely human-specific *E. coli* lineages (ST 69, ST 73, ST 95, ST 131) within Phylogroup B2 and D may be the most appropriate targets for the detection of human faecal contamination. Further, although these human-associated lineages are common in humans and when present in a host, numerically dominant, they are members of one of the least common phylogroups to be detected in water. Studies indicate that in the environment phylogroup B2 strains have a typical lifespan of 1-2 days (Berthe et al., 2013). The low frequency of phylogroup B2 strains in water samples is likely due to the fact they lose their culturability rapidly once released into water (Petit et al., 2017).

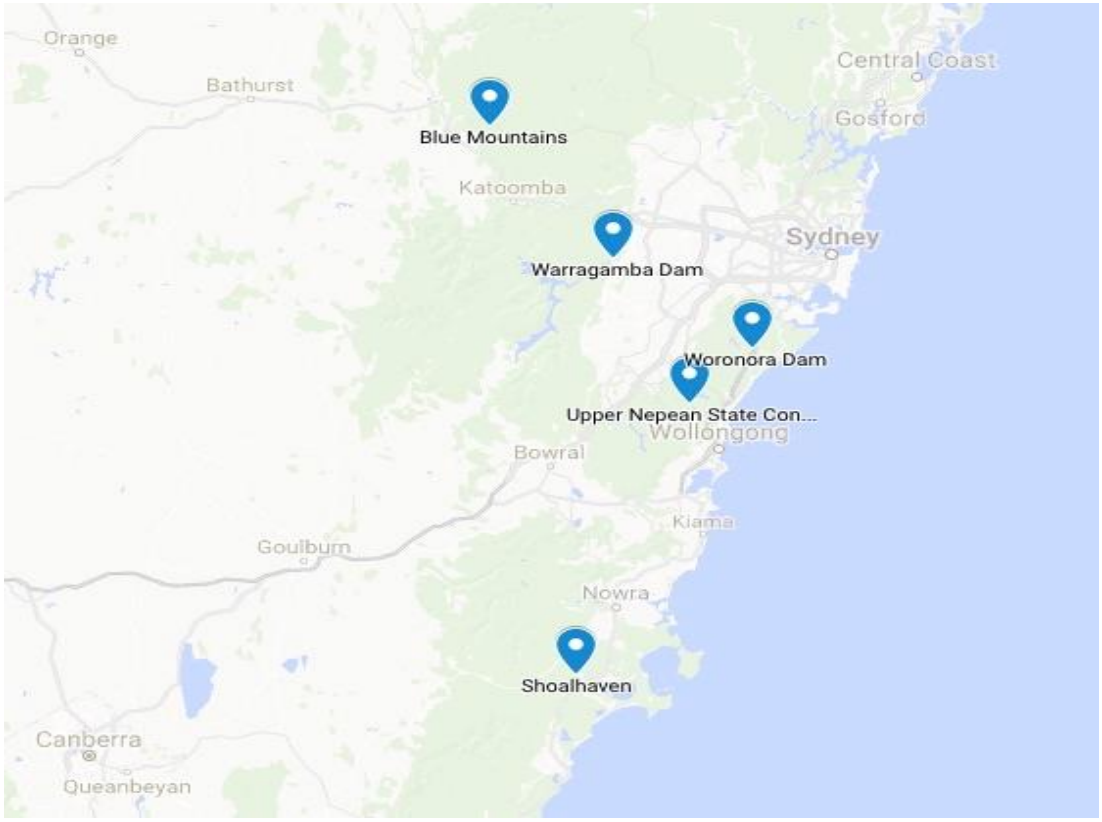
Thus, the purpose of this study is to characterise the frequency with which the human-associated lineages ST 69, ST 73, ST 95, ST 131 are detected in water bodies in eastern Australia.

Materials and Methods

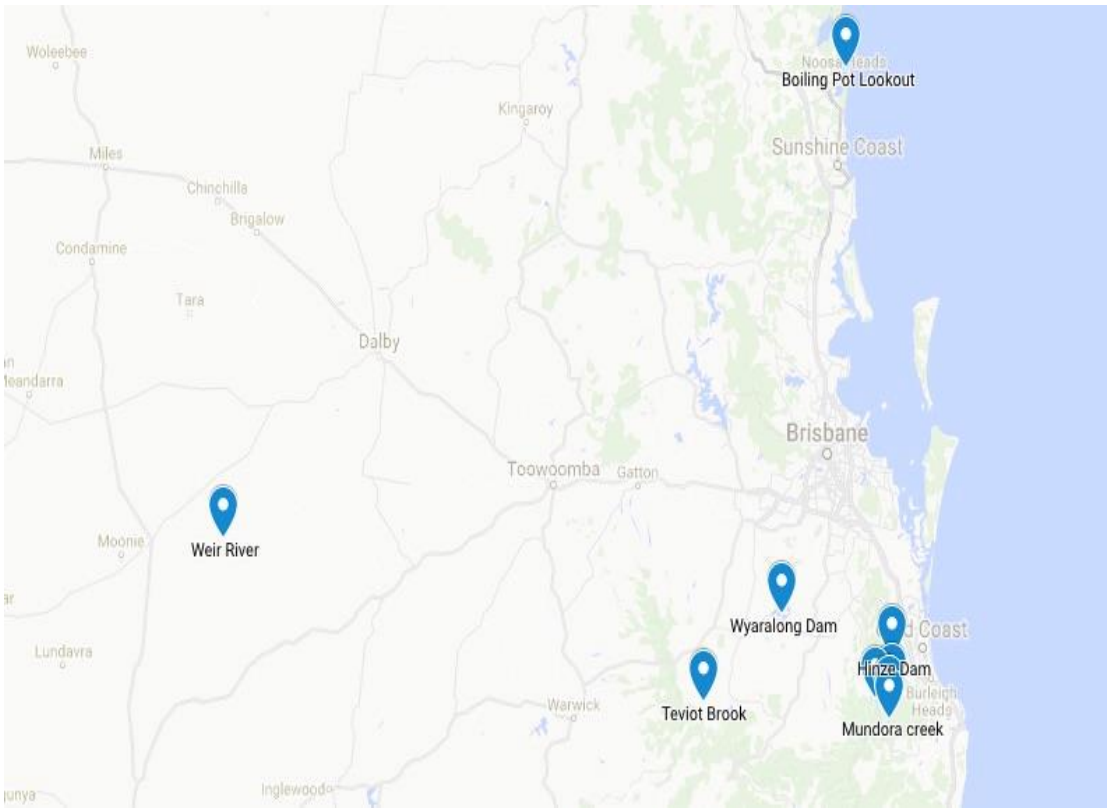
Water supply site description

To understand the frequency and distribution of human-associated *E. coli* lineages in a secondary environment, water samples were collected from two distinct geographic locations across Australia. The Water New South Wales (WaterNSW) water supply network provides drinking water for over 4.5 million people across Sydney and the Illawara, Blue Mountains, Southern Highlands, Goulburn and Shoalhaven regions. WaterNSW stores water in 11 major dams collected from rainfall in 5 vast catchments covering 16,000 square kilometers: Warragamba, Upper Nepean, Woronara, Shoalhaven and Blue Mountains. Water is transported from these dams via a complex system of rivers, weirs, pipes and canals into nine water filtration plants before reaching the household and business consumers in Sydney, Illawarra, the Shoalhaven, Blue Mountains and the Southern Highlands (https://www.waternsw.com.au/water-quality/education/learn/water-supply-system).

Seqwater delivers water to more than three million people in south-east Queensland. Surface water from natural open catchments across the Queensland region are stored in dams, weirs and off-stream storages such as Hinze Dam, Little Nerang Creek, Boiling Pot, Stafford Weir, Nixon Creek, Mudora Creek, Wyaralong Dam, Teviot Brook, Kuralboo Creek, Purling Brook, Boy-ull Creek. Seqwater operates 37 treatment plants across the region to treat the water from the dams and weirs through an inter connected water grid system to ensure high drinking water quality (http://www.seqwater.com.au/water-supply). It is then transported via a 600 km reversible flow pipeline network that provides drinking water to the regions of Sunshine Coast, Greater Brisbane, Redlands and Gold Coast. Both WaterNSW and Seqwater follow stringent processes and control measures to meet the ADWG in order to supply high quality drinking water.



Map 2.1 Water storage and catchment systems of WaterNSW, Sydney



Map 2.2 Water storage dams and weirs of Seqwater, Queensland

Strain selection

A total of 6089 *E. coli* isolates from WaterNSW, Sydney and 2167 isolates from Seqwater, Queensland were collected by both industries between December 2012 and December 2013 and sent for characterisation to the Gordon lab. In Gordon lab, all the isolates from these water samples were phylogrouped using the Clermont multiplex PCR for their phylogroup assignment (Clermont et al., 2013) and fingerprinted using ERIC and CGG primers (Versalovic et al., 1991; Adamus-Bialek et al., 2009) for their unique REP-profile. One isolate of each REP type per sample that belonged to phylogroup B2 and D were selected for further characterisation. From southeast Queensland there were 290 *E. coli* isolates belonging to phylogroup B2 and 88 isolates belonging to phylogroup D. There were 607 B2 and 486 D isolates in the collection from WaterNSW.

Identification of human-associated lineages

Until recently multi-locus sequence typing has been the 'gold standard' method for characterising *E. coli* (Wirth et al., 2006; Larsen et al., 2012). However, traditional MLST is costly and time-consuming. Recently two PCR based methods have been developed for specific detection of the human-associated clonal complexes. These are the B2 subtyping PCR and Doumith PCR. The B2 subtyping PCR detects the alleles associated with nine main human-associated lineages in the phylogroup B2 (Clermont et al., 2014). They are designated as I – X and correspond to the clonal complexes I (CC 131), II (CC 73), III (CC 127), IV (CC 141), V (CC 144), VI (CC 12), VII (CC 14), IX (CC 95), X (CC 372). The Doumith PCR detects genes specific to particular clonal lineages that are the most common host associated sequence types (STs) 69, 73, 95, 131 within phylogroups B2 and D (Doumith et al., 2015). These two PCR techniques were used in this study for the characterisation of phylogroup B2 and D strains at a clonal level and for determining the frequency of human-associated STs in the water samples.

B2 Subtyping PCR

All environmental B2 isolates with unique REP types were screened with allele-specific B2 subtyping PCR, that detects the nine-host associated sequence type lineages (Clermont et al., 2014). This PCR was done in 2 independent panels with multiplex PCR primers (Appendix- Table 1). The reaction was carried out using the 5x buffer (supplied

by BioLine), 2.5U Taq polymerase, 10uM of each primer with the final volume made up to 20ul. PCR was performed under the following conditions: 4 min denaturation at 94°C, 30 cycles for 5s at 94°C and 20s at 64°C and a final extension step of 5min at 72°C. PCR products were loaded on 2% agarose gel using TBE buffer. After electrophoresis, the gels were photographed using a UV transilluminator.

Doumith PCR

Doumith multiplex PCR (Doumith et al., 2015) was done on all B2 and D isolates with unique rep types. The PCR mixture contained the 5x buffer (supplied by BioLine), 2.5U Taq polymerase, 10uM each of eight primers (Appendix-Table 1) using 1.2ul of genomic DNA as the template. The final volume was made up to 20ul. PCR was performed in the following cycling conditions: An initial denaturation at 94°C for 3 min, 30 cycles of 94°C for 30 sec, 60°C for 30 sec, and 72°C for 30 sec and one final cycle of 72°C for 5 min. The amplified PCR products were run on 1.5% agarose and photographed under a UV transilluminator.

Doumith vs B2 subtyping

Both these PCR-based methods for detecting the main human associated B2 and D strains have their own limitations. For example, the sub-group I strains detected by B2-subtyping are not all part of CC 131 (Clermont et al., 2014), while not all strains assigned as CC 95 using the Doumith method are actually part of this complex (Gordon et al., 2017). Furthermore, although the ST 95 strains are frequently isolated from asymptomatic humans living in Australia, they are very rarely observed in Australian native vertebrates (Gordon et al., 2017). A comparison of 77 phylogroup B2 *E. coli* isolates with either whole genome sequence or MLST data reveals that when the two methods predict the same clonal complex then the prediction is correct 94% of the time and when the methods do not predict the same CC then the strain is unlikely to be a member of CCs 73, 95 or 131 (Table 2.1; Gordon et al., 2017; unpublished data).

Table 2.1 Percentage of strains correctly assigned based on B2 subtyping and Doumith PCR.

Strain CC Membership (MLST or WGS)	Both Doumith and B2 predict the CC	Only one of Doumith or B2 predict the CC
73 (n=35)	94.3% (33/35)	5.7% (2/35)
95 (n=30)	96.7% (29/30)	3.3% (1/30)
131 (n=12)	83% (10/12)	16.7% (2/12)

Consequently, as well as giving the results for the Doumith and B2 subtyping separately, the results for the two methods were also combined.

Results

Phylogroups B2 and D across sites

The isolation of *E. coli* isolates belonging to two main host associated phylogroups B2 and D varied across each site and sampling locations between the periods of December 2012 to December 2013. For the WaterNSW water samples, phylogroup B2 represented 12% and phylogroup D 9.6% of the REP types observed. For the Seqwater water samples, phylogroup B2 represented 19% and phylogroup D 5.8% of the REP types observed. The phylogroup B2 were comparatively under-represented in the WaterNSW samples, whilst phylogroup D were over-represented. Considering phylogroups B2 and D together, showed that the frequency of these phylogroups was similar in the Seqwater and WaterNSW samples; 25% versus 22% respectively.

B2 Subtyping PCR

The B2 subtyping PCR method was used to classify all the *E. coli* isolates belonging to phylogroup B2 from WaterNSW, Sydney and Seqwater, Queensland. Many of the group B2 strains remained unassigned to the nine-host associated B2-ST lineages as expected. Among the isolates that were classified into the host associated B2-ST lineages, ST 141 was frequently recovered followed by ST 127, and ST 12 was the least numerous in both locations (Table 2.2).

Table 2.2 Frequency of human-associated B2 lineages from WaterNSW and SeqWater as determined using B2 subtyping PCR.

Predicted STs	New South Wales N (% of all isolates n= 6089)	Queensland N (% of all isolates, n=2167)
12	9 (0.14%)	3 (0.13%)
127	56 (0.91%)	15 (0.69%)
131	33 (0.54%)	21 (0.96%)
14	12 (0.19%)	13 (0.59%)
141	85 (1.39%)	21 (0.96%)
144	15 (0.24%)	2 (0.09%)
372	21 (0.34%)	14 (0.64%)
73	23 (0.37%)	7 (0.32%)
95	9 (0.14%)	15 (0.69%)
UA*	344 (5.64%)	179 (8.26%)

* UA not one of the human associated STs.

Doumith PCR

A total of 1471 *E. coli* strains belonging to phylogroup B2 and D from WaterNSW, Sydney, and Seqwater, Queensland were assigned for the predominant sequence types (STs) 73, 95, 131 and 69 among the extra-intestinal pathogenic *Escherichia coli* (ExPEC) lineage using Doumith PCR.

Table 2.3 Frequency of human-associated B2 and D lineages from Water NSW and SeqWater as determined using the Doumith PCR.

Predicted STs	New South Wales N (% of all isolates n= 6089)	Queensland N (% of all isolates, n=2167)
3	114 (1.87%)	55 (2.53%)
95	56 (0.91%)	11 (0.51%)
31	17 (0.27%)	6 (0.28%)
69	38 (0.62%)	8 (0.37%)
UA	868 (14.25%)	298 (13.75%)

* UA not one of the human associated STs.

In the Doumith PCR, as expected most of the strains remained unassigned. ST 73 was relatively abundant among the predominant *E. coli* lineage STs across both sites and ST 131 was the least abundant (Table 2.3).

B2 Subtyping and Doumith comparision

Table 2.4 Correlation of results between the B2 subtyping PCR and Doumith PCR

Location & STs	B2 subtyping only	Doumith only	B2 subtyping and Doumith
WaterNSW, Sydney N (% of all isolates n= 6089)			
73	23 (0.37%)	114 (1.87%)	1 (0.01%)
95	9 (0.14%)	56 (0.91%)	4 (0.07%)
131	33 (0.54%)	17 (0.27%)	7 (0.11%)
Seqwater, Queensland N (% of all isolates, n=2167)			
73	7 (0.32%)	55 (2.53%)	0 (0%)
95	15 (0.69%)	11 (0.51%)	2 (0.09%)
131	21 (0.96%)	6 (0.28%)	4 (0.1%)

Combining the results of the B2 subtyping and Doumtith PCRs suggests that the human associated B2 STs 73, 95, and 131 typically represent less than 0.3% of all *E. coli* recovered from the water samples (Table 2.4).

Discussion

The aim of this study was to investigate the diversity of human associated *E. coli* phylogroups B2 and D in water, because these lineages predominate in humans as commensals, and human faecal contamination typically represents the greatest risk to water quality (EPA Victoria Guidelines, 2007; Sinton et al., 1998; McLellan and Eren, 2014). Consequently, the frequency of the most common human-associated *E. coli* lineages (B2: ST 73, ST 95, ST 131 and D ST 69) was assessed in water samples collected over a year from southeast Queensland (SeqWater) and eastern New South Wales (WaterNSW). The results of this survey demonstrated that these human-associated lineages represent, at most, typically less than 1% of the *E. coli* isolates recovered from water (Tables 2.2 and 2.3) are more likely to represent <0.1% of isolates (Table 2.4). This is despite the fact that strains belonging to phylogroups B2 and D typically represent >50% of the isolates recovered from humans living in Australia (Gordon et al., 2015; Dixit et al., 2018).

The rarity of these human associated sequence types is not unexpected. The extent to which humans have access to the WaterNSW and SeqWater catchments varies. The WaterNSW has strictly restricted areas for walking, picnicking and camping, while fishing, swimming, and boating are not permited in water storages in the Warragamba, Upper Nepean, Woronara, Blue Mountains catchment. The Shoalhaven system of catchments has a few recreational areas for direct public access such as Bendeela Recreational Area, Fitzroy Falls, Tallowa Dam and Wingecarribee Reservoir. Seqwater, on the other hand, allows public access to most of its water supply lakes and catchments. Only Hinze Dam is always restricted for powered boating and swimming. Despite the differing extent of human activities permitted in and around the catchments managed by SeqWater and WaterNSW the frequency of phylogroup B2 and D strains was similar, as was the frequency of the human-associated lineages.

The present study is in contrast with two other recent studies from France and Japan on rural catchments. In France, *E. coli* isolates were collected from water samples across four rivers in Normandy with similar hydrological conditions. Two of the rivers, the Selles and the Sebec were located near forests and dairy farms (Petit, 2017). In contrast, the rivers Tourville and Risle, which are located in the lower region of the catchments are

close to an urban centre. The authors demonstrated spatial changes in phylogroup abundance between rivers in the upper and lower regions of the catchment. Further, human-associated B2 lineages were not detected among isolates recovered from the rivers in the upper regions of the catchment, but were detected in the lower catchment region. Similarly, the frequency of ST 69 strains increased moving from the upper to the lower regions of the catchment. In Japan, characterisation of 531 *E. coli* isolates collected from the surface of Yamato River revealed 155 multi drug resistant isolates. Among the drug resistant isolates, 11.6% represented ST 95, 5.2% ST 131 and 5.2% were ST 69. The authors of the study concluded that the Yamato River was highly contaminated with clinically important *E. coli* lineages. They suggested that possible explanations for this contamination were the region's high population density, waste water treatment plants, animal husbandry facilities, and hospitals located close to or upstream of the sampling sites (Gomi et al., 2017). Compared to the above literatures, the results of this study infer a low likelihood of *E. coli* from human or animal sources in the catchments investigated and so, a low likelihood of antibiotic resistance load of significance.

Further, the results of this survey study challenge the two traditional views of looking at *E. coli*. Firstly, it challenges *E. coli* being used as a microbial source tracker based on an assumption that all *E. coli* are capable of transition from their primary environment to a secondary environment and therefore still reflecting the same clonal composition in both habitats. The study results instead indicate that the dominant STs from the human gut are rare in the water catchments tested. This suggests that the vast majority of other STs are likely to be useless as a microbial source tracker. Secondly, these results also indicate limitation of *E. coli* being considered and used as a faecal indicator organism, as majority (60%) of the isolates from phylogroup B2 and D remained unassigned to human-specific STs, suggesting that not all *E. coli* are faecally derived. However, the findings have implications that the dominant STs from human gut survive poorly in these waters and if they were found in large numbers and matched to an upstream source, they would likely represent recent human faecal origin and hence recent faecal contamination event. Finally, as most B2 and D isolates remain unassigned to human host associated ST lineages, the study indicates that the *E. coli* isolates from these water sources are not a result of recent faecal contamination and might even possibly be of a non-human origin.

References

Adamus-Bialek, W. et al., 2009. (CGG)4-Based PCR as a Novel Tool for Discrimination of Uropathogenic *Escherichia coli* Strains: Comparison with Enterobacterial Repetitive Intergenic Consensus-PCR. *Journal of Clinical Microbiology,* 47(12), p. 3937–3944.

Alm, E. W., Walk, S. T. & Gordon, D. M., 2011. The niche of *Escherichia coli*. In: S. T. Walk & P. C. H. Feng, eds. *Population Genetics of Bacteria.* Washington, DC, USA: ASM Press, pp. 107-123.

EPA Victoria Guidelines, Environment Protection (Scheduled Premises and Exemptions) Regulations 2007.

Bert, F. et al., 2010. Genetic Diversity and Virulence Profiles of *Escherichia coli* Isolates Causing Spontaneous Bacterial Peritonitis and Bacteremia in Patients with Cirrhosis. *Journal of Clinical Microbiology,* 48(8), p. 2709–2714.

Berthe, T. et al., 2013. Evidence for Coexistence of Distinct *Escherichia coli* Populations in Various Aquatic Environments and Their Survival in Estuary Water. *Applied and Environmental Microbiology,* 79(15), p. 4684–4693.

Bidet, P. et al., 2007. Detection and Identification by PCR of a Highly Virulent Phylogenetic Subgroup among Extraintestinal Pathogenic *Escherichia coli* B2 Strains. *Applied and Environmental Microbiology*, 73(7), p. 2373–2377.

Bingen, E. et al., 1998. Phylogenetic analysis of *Escherichia coli* strains causing neonatal meningitis suggests horizontal gene transfer from a predominant pool of highly virulent B2 group strains. *The Journal of Infectious Diseases,* 177(3), pp. 642-650.

Cabral, J. P., 2010. Water microbiology. Bacterial pathogens and water. *International Journal of Environmental Research and Public Health,* 7(10), pp. 3657-3703.

Clermont, O. et al., 2014. Development of an allele-specific PCR for *Escherichia coli* B2 sub-typing, a rapid and easy to perform substitute of multilocus sequence typing. *Journal of Microbiological Methods,* Volume 101, pp. 24-27.

Clermont, O., Christenson, J. K., Denamur, E. & Gordon, D. M., 2013. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports,* 5(1), pp. 58-65.

Clermont, O. et al., 2011. Characterization of the cryptic Escherichia lineages: rapid identification and prevalence. *Environmental Microbiology,* 13(9), pp. 2468-2477.

Day, M. J. et al., 2016. Population structure of *Escherichia coli* causing bacteraemia in the UK and Ireland between 2001 and 2010. *Journal of Antimicrobial Chemotherapy,* 71(8), pp. 2139-2142.

Desjardins, P. et al., 1995. Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment-length polymorphism. *Journal of Molecular Evolution,* 41(4), pp. 440-448.

Dixit , O. V. A., O'Brien, C. L., Pavli, P. & Gordon, D., 2018. Within-host evolution versus immigration as a determinant of *Escherichia coli* diversity in the human gastrointestinal tract. *Environmental Microbiology* , 20(3), pp. 993-1001.

Doumith, M. et al., 2015. Rapid identification of major *Escherichia coli* sequence types causing urinary tract and bloodstream infections. *Journal of Clinical Microbiology,* 53(1), pp. 160-166.

Edberg, S. C., Rice, E. W., Karlin, R. J. & Allen, M. J., 2000. *Escherichia coli*: the best biological drinking water indicator for public health protection. *Journal of Applied Microbiology,* 88(s1), pp. 106S-116S.

Escobar-Páramo, P. et al., 2006. Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates. *Environmental Microbiology,* 8(11), pp. 1975-1984.

Gibreel, T. M. et al., 2012. Population structure, virulence potential and antibiotic susceptibility of uropathogenic *Escherichia coli* from Northwest England. *The Journal of Antimicrobial Chemotheraphy,* 67(2), pp. 346-356.

Gomi, R. et al., 2017. Whole-genome analysis of antimicrobial-aesistant and extraintestinal pathogenic *Escherichia coli* in river water. *Applied and Environmental Microbiology,* 83(5), pp. e02703-16.

Gordon, D. M., 2013. The ecology of *Escherichia coli*. In: M. S. Donnenberg, ed. *Escherichia coli: Pathotypes and Principles of Pathogenesis.* Maryland, USA: Elsevier Inc., pp. 3-20.

Gordon, D. M. & Cowling, A., 2003. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology,* Volume 149, p. 3575–3586.

Gordon, D. M. et al., 2017. Fine-Scale Structure Analysis Shows Epidemic Patterns of Clonal Complex 95, a Cosmopolitan *Escherichia coli* Lineage Responsible for Extraintestinal Infectio. *mSphere,* 2(3), pp. e00168-17.

Gordon, D. M., O'Brien, C. L. & Pavli, P., 2015. *Escherichia coli* diversity in the lower intestinal tract of humans. *Environmental Microbiology Reports,* 7(4), pp. 642-648.

Gordon, D. M., Stern, S. E. & Collignon, P. J., 2005. Influence of the age and sex of human hosts on the distribution of *Escherichia coli* ECOR groups and virulence traits. *Microbiology,* 151(1), pp. 15-23.

Harwood, V. J. et al., 2014. Microbial source tracking markers for detection of fecal contamination in environmental waters: relationships between pathogens and human health outcomes. *FEMS Microbial Reviews,* Volume 38, pp. 1-40.

Herzer, P. J., Inouye, S., Inuye, M. & Whittam, T. S., 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. *Journal of Bacteriology,* 172(11), pp. 6175-6181.

Johnson, J. R., 2002. Evolution of Pathogenic *Escherichia coli*. In: M. S. Donnenberg, ed. *Escherichia coli-Virulence Mechanisms of a Versatile Pathogen.* Baltimore, Maryland: Elsevier BV, pp. 55-77.

Johnson, J. R., Delavari, P., Kuskowski, M. & Stell, A. L., 2001. Phylogenetic Distribution of Extraintestinal Virulence-Associated Traits in *Escherichia coli*. *The Journal of Infectious Diseases,* 183(1), pp. 78-88.

Kallonen, T. et al., 2017. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Research,* Volume 27, pp. 1-13.

Larsen, M. V. et al., 2012. Multilocus sequence typing of total-genome-sequenced bacteria. *Journal of Clinical Microbiology,* 50(4), pp. 1355-1361.

Le Gall, T. et al., 2007. Extraintestinal Virulence Is a Coincidental By-Product of Commensalism in B2 Phylogenetic Group *Escherichia coli* Strains. *Molecular Biology and Evolution,* 24(11), pp. 2373-2384.

Mahjoub-Messai, F. et al., 2011. *Escherichia coli* isolates causing bacteremia via gut translocation and urinary tract infection in young infants exhibit different virulence genotypes. *The Journal of Infectious Diseases,* 203(12), pp. 1844-1849.

Massot , M. et al., 2016. Phylogenetic, virulence and antibiotic resistance characteristics of commensal strain populations of *Escherichia coli* from community subjects in the Paris area in 2010 and evolution over 30 years. *Microbiology,* 162(4), pp. 642-650.

McLellan, S. L. & Eren, A. M., 2014. Discovering new indicators of fecal pollution. *Trends in Microbiology,* 22(12), pp. 697-706.

NHMRC-ARMCANZ, 1996. *Australian drinking water guidelines, 1996 / National Health and Medical Research Council, Agriculture and Resource Management Council of Australia and New Zealand.* Canberra.

NHMRC, N., 2011. *Australian Drinking Water Guidelines Paper 6 National Water Quality Management Strategy.* Canberra: National Health and Medical Research Council, National Resource Management Ministerial Council, Commonwealth of Australia.

Nowrouzian, F. L., Adlerberth, I. & Wold, A. E., 2006. Enhanced persistence in the colonic microbiota of *Escherichia coli* strains belonging to phylogenetic group B2: role

of virulence factors and adherence to colonic cells. *Microbes and Infection,* 8(3), pp. 834-840.

Petit, F. et al., 2017. Change in the Structure of *Escherichia coli* Population and the Pattern of Virulence Genes along a Rural Aquatic Continuum. *Frontiers in Microbiology,* 8(609), p. doi: 10.3389/fmicb.2017.00609.

Regli, S., Rose, J. B., Haas, C. N. & Gerba, C. P., 1991. Modelling risk for pathogens in drinking water. *Journal of the American Water Works Association,* 83(11), pp. 76-84.

Selander, R. K. & Levin, B. R., 1980. Genetic diversity and structure in *Escherichia coli* populations. *Science,* 210(4469), pp. 545-547.

Seqwater, Available at: http://www.seqwater.com.au/water-supply [Accessed 15 May 2017].

Sinton, L. W., Finlay, R. K. & Hannah, D. J., 1998. Distinguishing human from animal faecal contamination in water: A review. *New Zealand Journal of Marine and Freshwater Research,* 32(2), pp. 323-348.

Smati, M. et al., 2015. Quantitative analysis of commensal *Escherichia coli* populations reveals host-specific enterotypes at the intra-species level. *Microbiology,* 4(4), pp. 604-615.

Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E., 2010. The population genetics of commensal *Escherichia coli*. *Nature Reviews Microbiology,* Volume 8, pp. 207-217.

Versalovic, J., Koeuth, T. & Lupski, J. R., 1991. Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. *Nucleic Acid Research,* 19(24), p. 6823–6831.

WaterNSW, Available at: https://www.waternsw.com.au/water-quality/education/learn/water-supply-system [Accessed 20 Feb 2017].

WHO, 1993. *Guidelines for Drinking Water Quality.* Second ed. World Health Organization, Geneva: Volume 2 Health criteria and other supporting information.

Wirth, T. et al., 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Molecular Microbiology,* 60(5), pp. 1136-1151.

CHAPTER 3

Pan genome comparison of *E. coli* phylogroup B2 isolates from humans and native Australian vertebrates

Introduction

Escherichia coli is a diverse species living as a commensal in the large intestine of humans and other warm-blooded animals, as well as in soil and aquatic environments (Alteri and Mobley, 2012; Katouli, 2010; van Elsas et al., 2011; Walk et al., 2007). The species is predominantly clonal and this allows them to be delineated into phylogroups (Tenaillon et al., 2010). Currently, *E. coli* is classified into eight phylogroups A, B1, B2, C, D, E, F and cryptic clade I (Clermont et al., 2013; Clermont et al. 2015). Each phylogroup varies in its life history characteristics, ecological niches, phenotypic characteristics and ability to cause disease (Bergthorssonm and Ochman, 1998; Johnson et al., 2001; Gordon and Cowling, 2003; Anastasi et al., 2010; Carlos et al., 2010).

With recent advances in whole genome sequencing techniques, our understanding of the genetic diversity exhibited within *E. coli* is also rapidly increasing. Thus far, it is understood that the *E. coli* genome consisting of 4,000 to 5,500 genes individually is collectively composed of the conserved core of approximately 2000 genes that is common to most of the *E. coli* population, and a flexible gene pool of currently > 90 000 genes (Land et al., 2015). This large variable gene pool is due to the combination of the gain of genes by horizontal gene transfer and the loss of genes through deletion. The flexible gene pool constitutes the genetic information that provides *E. coli* the ability to cope with broad range of habitats and various ecological conditions (Touchon et al., 2009; Oh et al., 2012; Bielaszewska et al., 2007; de Muinck et al., 2013; Blount, 2015).

For decades, the primary habitat of *E. coli* was believed to be the lower intestinal tract of warm-blooded animals. The concentration of *E. coli* ranges from 10^7- 10^9 per gram in human faeces to 10^4- 10^6 per gram in domestic animal faeces (Tenaillon et al., 2010). Natural environments such as soil and water were considered to be the secondary habitat of *E. coli* as their presence was thought to be solely a consequence of faecal inputs (Berthe et al., 2013; Edberg et al., 2000). However, recent findings indicate that a substantial population of *E. coli* not only survives but actually multiplies in these secondary habitats

(Walk et al., 2009; Ishii et al., 2006; Luo et al., 2010) and these isolates have been called naturalised and free-living *E. coli*. Naturalised or environmentally adapted *E. coli* isolates are strains that may have originated from faecal contamination but over time became stress tolerant and environmentally adapted, losing virulence due to mutations resulting in niche-specific adaptation (Walk et al., 2007; Chiang et al., 2011). Free living environmental isolates are strains whose persistence in secondary habitat is independent of any faecal inputs (Power et al., 2005).

E. coli is used as the principal indicator of faecal contamination (faecal indicator bacteria (FIB)) to test waterways across the globe (USEPA, 1986; Ashbolt et al, 2001; Australian Drinking Water Guidelines 6, 2011). *E. coli* is used as FIB with the assumption that they are present at a high number in faeces of humans and other warm-blooded animals, and considered to have poor survival in water. It is considered as a transient resident in secondary habitat (Odonkor and Ampofo, 2013). Yet, numerous recent studies report on *E. coli* isolates that not only survive for a long period and replicate in external environments, but that are also distinct from the *E. coli* isolated from humans (Power et al., 2005; Byappanahalli et al., 2006; Walk et al., 2007; Ishii and Sadowsky, 2008; Vignaroli et al., 2015). Furthermore, *E. coli* is also used as a FIB due to the rapid and simple laboratory detection methods available for this species (Ashbolt et al., 2001). Historically, culture-based methods and enzyme based colorimetric kit methods are used for identification of *E. coli* from environmental samples (Jang et al., 2017). The major drawback of these methods is that they fail to differentiate the true human faecal isolates from isolates present in water due to animal waste, and naturalised or environmentally adapted isolates that persist in the secondary habitat for a long period of time (Field and Samadpour, 2007; Harwood et al., 2014; Gomi et al., 2014; Ahmed et al., 2015). Failure to identify the source of contamination could lead to inaccurate interpretation of microbiological monitoring data, therefore resulting in over or under estimating the public health risk (Ferguson et al., 1996; Wade et al., 2008; Harwood et al., 2014). Hence, in order to detect and assign the *E. coli* loading in water to its correct source, water industries, sanitary engineers and government authorities need a probe/ genetic marker that could detect and differentiate *E. coli*'s source of origin, that is human, animal or naturalised.

Members of the *E. coli* phylogroups vary in their ecological niche, life history characteristics and propensity to cause disease. Phylogroup A and B1 are considered to be better survivors in external environment, while phylogroup B2 and D strains are considered survive poorly in these habitats (Quero et al., 2015). Phylogroup B2 strains and D strains are more likely to be detected in endothermic vertebrates as compared to ectotherms, while B2 strains are more likely to be detected in mammals compared to birds. Among mammals, B2 strains are isolated more often from species with a hindgut fermentation chamber compared to species lacking a caecum (Gordon & Cowling 2003). In industrialised countries like Australia, phylogroup B2 strains are the phylogroup most frequently isolated from human faeces, blood and urine (Picard et al., 1999; Gordon et al., 2005; Walk et al., 2007; Touchon et al., 2009; Gordon et al., 2017). They are thought to be the most specialised and host adapted strains, persisting in individual humans longer than strains of most other phylogroups (Nowrouzian et al., 2005; Clermont et al., 2008; Smati et al., 2013). Interestingly, strains of this phylogroup are far more likely to cause extra-intestinal diseases compared to strains belonging to other phylogroups (Johnson, 2002; Johnson and Russo, 2002; Day et al., 2016). They are also frequently detected in livestock, poultry, and companion animals, as well as a number of wild species (Gordon and Cowling, 2003; Clermont et al., 2011; Blyton et al., 2013; Coura et al., 2015; Alonso et al., 2017).

The primary aim of the research described in this chapter is to discover if there are any genetic markers that may distinguish human *E. coli* isolates from animal derived *E. coli*, and investigate if there might be a subpopulation of B2 strains naturalised to the aquatic environment.

Materials and Methods

Strain selection

The Gordon Laboratory has a large collection of *E. coli* isolates from variety of host sources such as Australian vertebrates, including humans and animals, and from various water sources (Gordon and Cowling, 2003; Gordon et al., 2005; Power et al., 2005; Blyton et al., 2015; Gordon et al., 2015; Blyton and Gordon, 2017; Dixit et al., 2018). Whole genome sequencing has been carried out on a subset of this collection. To select the strains for this study, a whole genome single nucleotide polymorphism (SNP) based

phylogenetic tree was constructed from all the whole genome sequenced strains belonging to phylogroup B2. From each node pairs were chosen with the following criteria: one isolate recovered from a water sample and one isolate taken from a host (Figure 3.1). Water isolates were sourced from drinking water catchments across Sydney and southeast Queensland. Host strains were derived from either humans or native vertebrates.

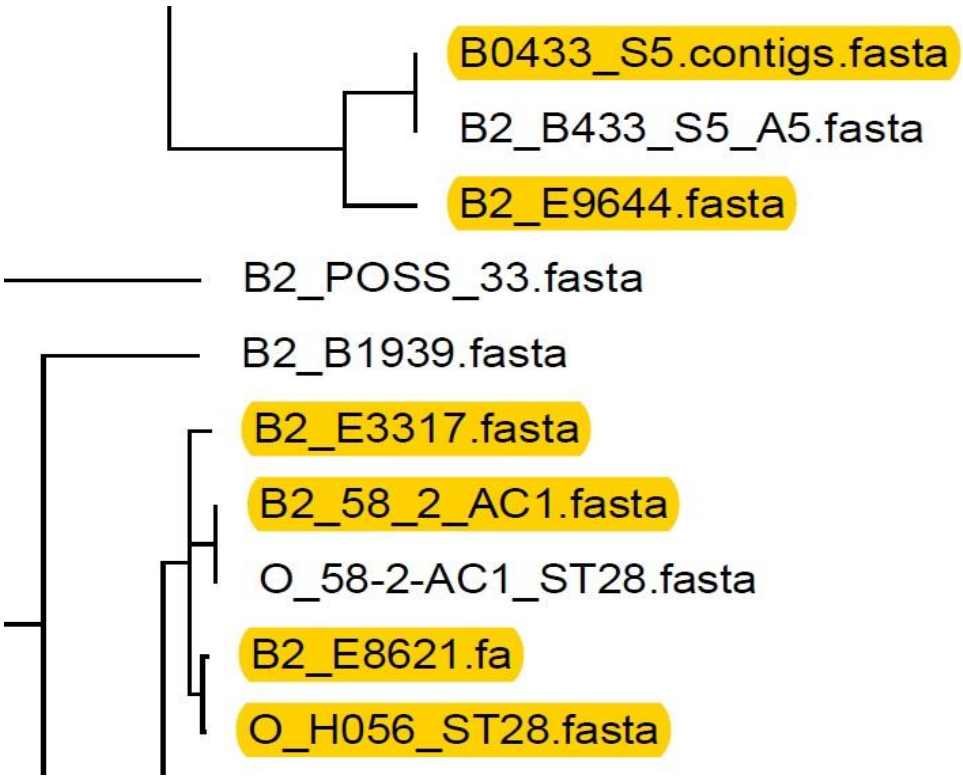


Fig. 3.1 Phylogenetic tree illustrating the phylogroup B2 strains chosen for this study. For each node one environmental water isolate was selected (strain name starting with an E) and one isolate from a host of either human or native vertebrate.

A total of 75 strains were selected and these included 36 water environmental strains and host strains of 12 human isolates and 27 native vertebrate isolates. A total of 42 STs were represented among the 75 isolates.

Table 3.1 Table summarises the number of strains in each of the STs

ST	Count of isolates
12	1
28	4
95	8
126	3
127	1
135	3
136	1
355	1
372	2
491	2
569	2
589	2
636	1
681	4
1257	1
1386	2
1619	1
1800	1
1858	2
1873	2
1894	1
1899	1
1925	2
2474	1
2622	1
2800	2
3290	1
3291	1
3304	1
3306	1
3307	6

ST	Count of isolates
3646	1
3672	1
5430	1
6165	1
6947	1
6948	1
6949	3
6950	1
6952	1
6998	1
Unknown ST	1
Grand Total	**75**

Pangenome analysis

Whole genome sequencing was performed using Illumina Nextera–library preparation kits and sequenced with the Illumina MiSeq platform using V3 chemistry (2x300 paired end reads) and reads of all the strains were assembled using A5miseq assembly software (Coli et al., 2015). Assemblies were annotated using the software Prokka (Seemann, 2014). Pan genome analysis was undertaken using the Roary (Page et al., 2015). Scoary was used to determine the frequency of each gene with respect to the source of the isolate (Brynildsrud et. al., 2016). Roary estimated there were 18367 genes in the pan genome; genes present in > 90% of the strains or < 10% of the strains were eliminated from the data. This elimination of genes was done to study potentially informative genes of significance as > 90% and <10% are the most common and most rare genes that may cause statistical insignificance to the data. Hence, this reduced the data set to 2295 genes.

Statistical analysis and gene exploration

The presence/absence matrix of the 2295 genes in the 75 B2 isolates was used to visualise the relationships among the strains as determined by their variable gene content using a Principal Coordinate Analysis (PCoA) based on Jaccard similarity coefficients. Further analysis, such as one-way Analysis of similarities (ANOSIM) and Permutational

Multivariate Analysis of Variance (PERMANOVA) using Jaccard similarity coefficients were used to determine if isolate source (water *vs* host) explained any of the observed variation. Effects of variation were considered to be significant only when the probability values were less than 0.05. The PCoA, ANOSIM and PERMANOVA were all computed using the PAST3 software (Hammer et al, 2001). The MLST, antimicrobial resistance, plasmids, virulence factors and serotypes of each isolate was identified using Center for genomic epidemiology website (http://www.genomicepidemiology.org). Annotation of the genes of importance were done by Genoscope Microscope (http://www.genoscope.cns.fr/agc/microscope/search/blast.php?) and Microbial Nucleotide BLAST search (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=MicrobialGenomes). Further details of the analyses undertaken will be described in the results and discussion sections.

Results

The pangenome analysis was based on 75 *E. coli* isolates belonging to phylogroup B2, and consisted of 36 water isolates, 39 host isolates representing 42 different STs. The results from the PCoA analysis using the presence/absence matrix of genes belonging to humans, native vertebrates and water isolates primarily indicated variation in the dataset based on the strains source of isolation (Figure 3.2). Isolates from native vertebrates were mainly concentrated in the upper right quadrant of the PCO plot, while isolates from humans largely clustered in the lower left quadrant. Water isolates were scattered across the PCO plot.

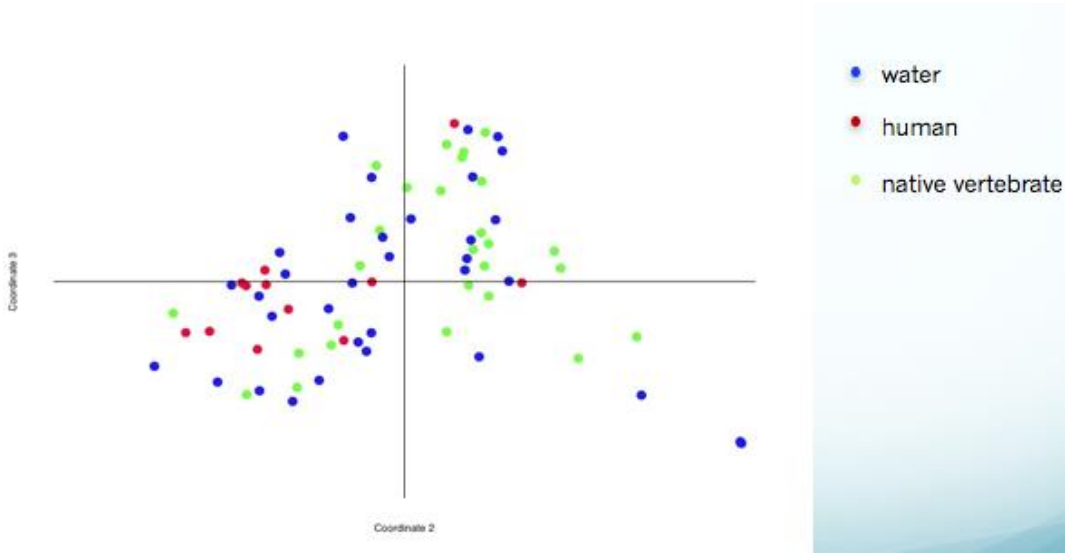


Fig. 3.2 Scatter plot of initial PCoA analysis

One-way ANOSIM and PERMANOVA tests with 9999 permutations revealed that while human and native vertebrates had average distinct variable gene profiles, isolates from water were not consistently distinct (P>0.05) from host isolates (Table 3.2).

Table 3.2 One-way ANOSIM and One-way PERMANOVA *p* values comparison with respect to source

Sources	**One-way ANOSIM**		**One-way PERMANOVA**	
	R value	***p* value**	**F value**	***p* value**
Water vs. Human	0.1184	0.0386	1.355	0.1121
Water vs. Native vertebrate	-0.0258	0.8319	0.977	0.4806
Human vs. Native vertebrate	0.3019	0.0003	2.179	0.0013

Inspection of the PCoA revealed two distinct clusters of strains; one over-represented by isolates from humans (cluster 1) and the other over-represented by isolates from native vertebrates (cluster 2) (Figure 3.3).

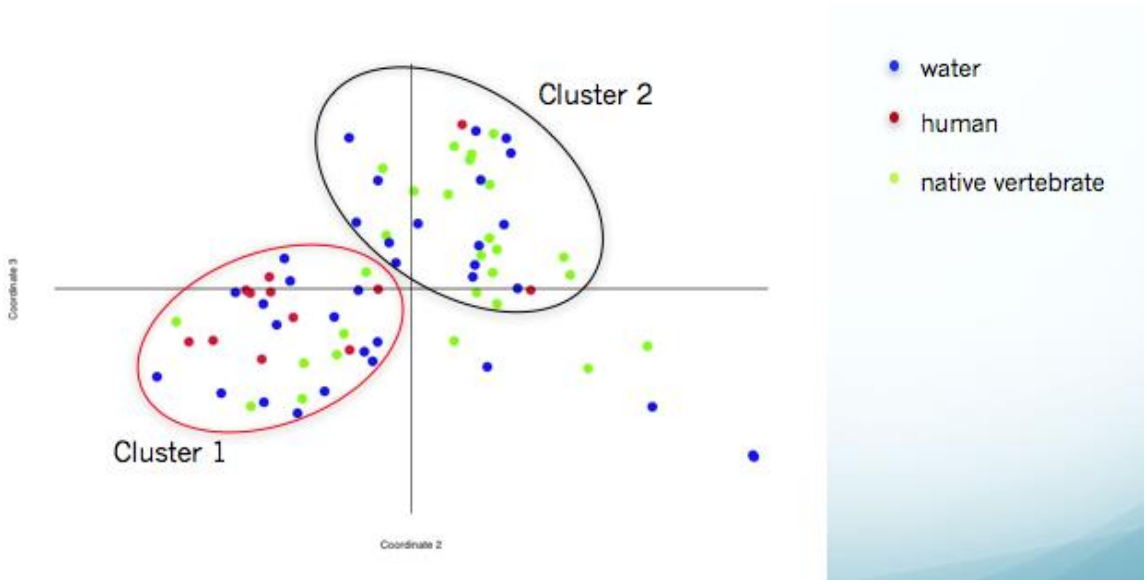


Fig. 3.3 Scatter plot of PCoA analysis representing Cluster 1 and Cluster 2

Of the 75 isolates, 43% of them were in Cluster 1, 45% in Cluster 2 and 12% could not be assigned to a cluster. Isolates from humans and native vertebrate animals were clearly non-randomly distributed between the two clusters (Contingency table analysis, X2=9.5; P>X2=0.009). Cluster 2 is dominated by isolates from native animals, while isolates from humans are more common in Cluster 1 than are animal isolates. Isolates from water are equally represent in both clusters (Figure 3.4).

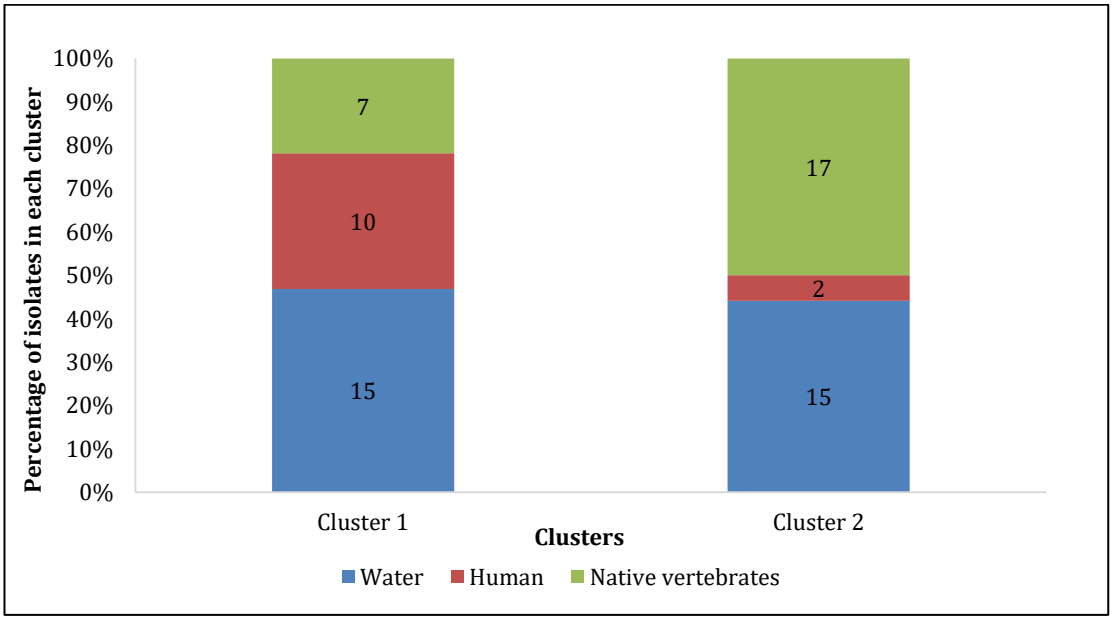


Fig. 3.4 Stacked bar chart summarising the proportion of isolates from each source in the Cluster 1 and Cluster 2

Considering the isolates in Cluster 1 and Cluster 2 separately, ANOSIM and PERMANOVA analyses revealed that source of isolation (water vs host or native animal vs human) did not explain a significant amount of the variation observed within a cluster (data not presented).

Genetic exploration

Since, human and native vertebrate isolates were non-randomly distributed among the 2 PCO clusters, the decision was made to analyse the variable gene content of isolates belonging to each of the PCO clusters. Out of the 2295 genes with frequency > 10% and < 90%, Cluster 1 with 32 isolates had variable genome of 2249 genes and Cluster 2 with 34 isolates had 2198 genes in its variable genome. The genes of significance were defined as the genes that were over-represented in one cluster but under-represented in the other; that is genes occurring in more than 2/3rd of isolates in one cluster but less than 1/3rd in the other cluster (Appendix B). The genes considered to be significant together with their frequency in each cluster are presented in Tables 3.3 and 3.4. The function of genes and their association with respective clusters are further detailed in discussion of this chapter.

Table 3.3 Genes over-represented (≈>66%) in strains from Cluster 1 and under-represented (≈<33%) in strains belonging to Cluster 2.

Gene	Annotation	Cluster 1 (%)	Cluster 2 (%)
yddA	ABC transporter ATP binding protein	72%	29%
yddB	putative porin protein	72%	29%
repB	RepFIB replication protein A	72%	26%
group_969	hypothetical protein	72%	21%
cbtA	*CbtA-CbeA* toxin-antitoxin system	72%	9%
gntP	gluconate / fructuronate transporter	69%	32%
mokB	peptide regulating *hokB* expression	69%	29%

Gene	Annotation	Cluster 1 (%)	Cluster 2 (%)
xerD_2	site-specific recombinase	69%	26%
group_2288	putative transporter subunit	69%	21%
group_3859	hypothetical protein	69%	18%
ydcM_2	putative transposase	69%	9%
yeeT_3	CP4-44 prophage; predicted protein	69%	9%
etk	tyrosine kinase	69%	0%
etp	phosphotyrosine-protein phosphatase	69%	0%
gfcA	inner membrane protein	69%	0%
gfcB	putative outer membrane lipoprotein	69%	0%
gfcD	putative lipoprotein	69%	0%
group_4796	hypothetical protein	69%	0%
group_6997	Putative exopolysaccharide export protein	69%	0%

Table 3.4 Genes over-represented (≈>66%) in strains from Cluster 2 and under-represented (≈<33%) in strains belonging to Cluster 1.

Gene	Annotation	Cluster 1 (%)	Cluster 2 (%)
appA	Acid phosphatase	38%	100%
gfcE	exopolysaccharide export protein	25%	94%
group_1511	hypothetical protein	25%	94%
group_8015	hypothetical protein	25%	94%
kdsB_2	3-deoxy-D-manno-octulosonate cytidylyltransferase	25%	94%
kdsD_3	D-arabinose 5-phosphate isomerase	25%	94%
yncG	putative glutathione s-transferase	31%	94%
group_887	hypothetical protein	22%	91%
group_5719	hypothetical protein	25%	91%
kpsM	Polysialic acid transport protein *KpsM*	25%	88%

Gene	Annotation	Cluster 1 (%)	Cluster 2 (%)
group_4152	hypothetical protein	31%	88%
group_7909	hypothetical protein	31%	88%
yvqK	Cob(I)yrinic acid a,c-diamide adenosyltransferase	31%	88%
group_3084	hypothetical protein	31%	82%
rsxC_2	member of *SoxR*-reducing complex	31%	82%
group_1717	hypothetical protein	25%	76%
group_1268	hypothetical protein	28%	74%
ygcG_3	putative protein	31%	74%
bglG_1	*BglG* transcriptional antiterminator	28%	71%
agaR_3	DNA-binding transcriptional repressor	31%	71%
fabG_3	3-oxo-acylreductase	31%	71%
gatA	galactitol PTS permease - *GatA* subunit	31%	71%
gatC_3	galactitol PTS permease - *GatC* subunit	31%	71%
group_4164	hypothetical protein	31%	71%
rpiB_2	allose-6-phosphate isomerase / ribose-5-phosphate isomerase B monomer	31%	71%
sgcB_2	putative enzyme IIB component of PTS	31%	71%
yjhU	predicted DNA-binding transcriptional regulator	31%	71%
group_3015	2,3-diketo-L-gulonate:Na+ symporter - membrane subunit	28%	68%
group_883	hypothetical protein	28%	68%
hxlB_1	3-hexulose-6-phosphate isomerase	31%	68%
rpe_2	ribulose-5-phosphate 3-epimerase	31%	68%

Distribution of *eae* gene in Cluster 1 and Cluster 2

The eae gene is considered to represent a virulence factor for several pathovars capable of causing intestinal disease, including enterohaemorrhagic *E. coli* (EHEC), enteropathogenic *E. coli* (EPEC), and attaching and effacing *E. coli* (AEEC) (Donnenberg et al., 1993 a & b; Cid et al., 2001; Kobayashi et al., 2003; Blanco et al., 2006). The *eae* gene is part of the locus of enterocyte effacement (LEE) in these pathovars and is the determinant for the production of intimin proteins that facilitates attaching and effacing (A/E) lesions in host epithelial cells during infection. Pathogenic strains with this gene are also thought to be able to survive and grow in natural environments (Jang et al., 2017). In the analysis done for this thesis, it was observed that the *eae* virulence factor gene was present only in Cluster 1 and not in Cluster 2 isolates (Appendix C). Within Cluster 1 it was present in three of the humans (9%) and three water (9%) isolates. The six strains positive for *eae* virulence gene also had the *etk, etp, gfc (A, B, D),* group 4796 and group 6997 genes.

Discussion

The present study investigated the similarities and differences of *E. coli*'s genetic composition using isolates from human and native vertebrate hosts, and from water environmental isolates. The PCoA analysis showed a clear separation of the 75 isolates belonging to phylogroup B2 into two separate clusters, with the isolates from humans over-representing one cluster, and isolates from native vertebrates over-representing the other. Of the 19 genes over-represented among Cluster 1 isolates (Table 3.3) 5 genes were associated with metabolic regulation, 8 genes with virulence such as toxins and capsule formation, and 6 genes encoded for hypothetical proteins. The genes *mokB* and *cbtA* are responsible for the Toxin and Antitoxin (TA) system in *E. coli*. Over-expression of *mokB* has been shown to kill cells or induce cells to enter a persistent state in which they transiently survive antibiotic exposure (Gerdes, 2016), while over-expression of *cbtA* decreases growth and colony formation (Heller et al., 2017). The gene *gntP* had transport function for gluconate uptake (Klemm et al., 1996) and *xerD* aids in site specific DNA recombination (Grainge and Sherratt, 1999). The genes *yddA* and *yddB* are required for optimal growth of *E. coli* at 37 °C (Serina et al., 2004). They are also upregulated in

UPEC isolates during the urinary tract infection process in mammalian hosts (Subashchandrabose et al., 2013).

The balance of the genes, *etk, etp, gfc (A, B, D)* and two hypothetical proteins (group 4796 and group 6997) were present only in Cluster 1 isolates and not in Cluster 2. The genes (*etk, etp, gfcABD*) are essential for Group 4 capsule (G4C) formation. G4C capsule is otherwise known as O-antigen capsule, which plays a significant role in attachment during the infection process (Peleg et al., 2005; Whitfield, 2006; Thomassin et al., 2013). The pathogenic intestinal isolates EPEC and EHEC are known to produce the G4C before infection by attachment onto intestinal epithelial cells (Sathiyamoorthy et al., 2011; Thomassin et al., 2013). It is also known that EPEC and EHEC cause infection by intestinal lesions known as attaching and effacing (A/E) lesions. These lesions are produced by intimins, encoded by *eae* virulence gene that is located at the pathogenicity island, locus of enterocyte effacement (LEE) (Kaper et al., 2004; Hazen et al., 2013). Deletion of *eae* gene has shown reduced pathogenicity of EPEC isolates in human volunteers (Donnenberg et al., 1993a). The gene *eae* is also one of the frequently detected virulence genes in *E. coli* from environmental samples (Ishii et al., 2014). A recent review by Jang et al., 2017 suggests that pathogenic *E. coli* isolates positive for *eae* gene can grow in natural environments. In this study, the virulent *eae* gene was detected in a total of 6 isolates (3 human and 3 water) belonging to Cluster 1. These isolates also contained the G4C producing capsule genes, suggesting these strains could be pathogenic.

In Cluster 2, out of the 31 genes of significance (Table 3.4), only one gene, *kpsM* was associated with virulence. The gene *kpsM* is responsible for Group 2 Capsule (G2C) formation in ExPEC strains (Whitfield and Roberts, 2002; Johnson and O'Bryan, 2004). It was present in 30 of the cluster 2 isolates and 8 Cluster 1 isolates. 17 other genes were related to sugar transport, metabolic regulation and transcription regulation, and 13 genes were hypothetical in function. One of the metabolic genes, the gene *appA* was present in all 34 isolates within the Cluster 2 (100%) and in 38% (5 water, 5 human and 2 native vertebrate) of Cluster 1 isolates. This gene encodes for acid phosphatase/ Phytase enzyme activity in *E. coli* (Greiner et al., 1993; Rodriguez et al., 1999; Yoon et al., 2011). The phytase enzyme hydrolyses the organic phosphate of phytic acid into inorganic phosphate. Phytic acid in organic form is the major storage form of phosphorous in cereals, legumes, oil seeds and nuts (Mullaney et al., 2000). Ruminant animals and humans with vegetarian

or vegan diets are known to have *E. coli* in their gut that produce phytase to hydrolyse the organic phosphate of phytic acid into easily digestible inorganic phosphate (Markiewicz et al., 2013). In this study, 1 native vertebrate isolate of Cluster 1 and 2 of the 17 native vertebrate isolates of Cluster 2 positive for *appA* gene were known to be from a herbivorous source. Also, *appA* is regulated by the stress response gene *rpo*S and it is induced when cells enter stationary phase (Touati et al., 1987; Lange and Hagge-Aronis, 1991). The presence of these two genes *appA* and *kpsM* is highly correlated in Cluster 2, but the importance of this observation in cluster 2 remains unclear and needs further research.

Understanding the genetic variation of *E. coli* isolates from various ecological conditions or hosts is critically important for its use as a FIB. This is the first study reporting of genetic variation within phylogroup B2 isolates from various sources such as humans, animals, water environment. This study aimed at discovering if there are any genetic markers to distinguish human *E. coli* isolates from animal derived *E. coli*, and investigate if there might be a subpopulation of B2 strains naturalised to the aquatic environment. The results of this study are significant in two respects in line with the aim. First, the findings suggest that the water isolates from human and native vertebrates could be distinguished using potential genetic markers. Water isolates with the *eae* and G4C capsular genes were more likely to come from a human source while those with *appA* were more likely to be from a native vertebrate source. Second, these data also suggests that there may not be naturalised isolates of phylogroup B2 in water. Overall, this study concludes that since *E. coli* is genetically diverse, not all *E. coli* are indicators of human faecal contamination and may not be used as the best generalized human faecal indicator.

References

Alonso, C. A. et al., 2017. High frequency of B2 phylogroup among non-clonally related fecal *Escherichia coli* isolates from wild boars, including the lineage ST131. *FEMS Microbiology Ecology*, 93(3).

Alteri, C. J. & Mobley, H. L., 2012. *Escherichia coli* physiology and metabolism dictates adaptation to diverse host microenvironments. *Current Opinion in Microbiology*, 15(1), pp. 3-9.

Anastasi, E. M. et al., 2010. Prevalence and Persistence of *Escherichia coli* Strains with Uropathogenic Virulence Characteristics in Sewage Treatment Plants. *Applied and Environmental Microbiology*, 76(17), pp. 5882-5886.

Ashbolt, N. J., Grabow, W. O. K. & Snozzi, M., 2001. Indicators of microbial water quality. In: L. F. a. J. Bartram, ed. *World Health Organization (WHO). Water Quality: Guidelines, Standards and Health.* London, UK: IWA Publishing, p. ISBN: 1 900222 28 0.

Bergthorsson, U. & Ochman, H., 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Molecular Biology and Evolution*, 15(1), pp. 6-16.

Berthe, T. et al., 2013. Evidence for Coexistence of Distinct *Escherichia coli* Populations in Various Aquatic Environments and Their Survival in Estuary Water. *Applied and Environmental Microbiology*, 79(15), p. 4684–4693.

Bielaszewska, M. et al., 2007. Shiga Toxin-Mediated Hemolytic Uremic Syndrome: Time to Change the Diagnostic Paradigm?. *PLoS ONE*, 2(10), p. e1024.

Blanco, M. et al., 2006. Typing of intimin (eae) genes from enteropathogenic *Escherichia coli* (EPEC) isolated from children with diarrhoea in Montevideo, Uruguay: identification of two novel intimin variants (muB and xiR/beta2B). *Journal of Medical Microbiology*, 55(9), pp. 1165-1174.

Blount, Z. D., 2015. The unexhausted potential of *E. coli*. *eLife*, Issue 4, p. e05826.

Blyton, M. D. J. et al., 2014. Not all types of host contacts are equal when it comes to*E. coli* transmission. *Ecology Letters* , Volume 17, pp. 970-978.

Blyton, M. D. J. & Gordon, D. M., 2017. Genetic Attributes of *E. coli* Isolates from Chlorinated Drinking Water. *PLoS ONE,* 12(1), p. e0169445.

Blyton, M. D. et al., 2015. Genetic Structure and Antimicrobial Resistance of *Escherichia coli* and Cryptic Clades in Birds with Diverse Human Associations. *Applied and Environmental Microbiology,* 81(15), p. 5123–5133.

Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V., 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome biology,* 17(1), pp. 238-247.

Byappanahalli , M. N. et al., 2006. Population structure, persistence, and seasonality of autochthonous *Escherichia coli* in temperate, coastal forest soil from a Great Lakes watershed. *Environmental Microbiology,* 8(3), pp. 504-513.

Carlos, C. et al., 2010. *Escherichia coli* phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. *BMC Microbiology,* 10(161).

Centre for genomic epidemiology. Available at: http://www.genomicepidemiology.org [Accessed 24 September 2017].

Chiang, S. M., Dong, T., Edge, T. A. & Schellhorn, H. E., 2011. Phenotypic Diversity Caused by Differential RpoS Activity among Environmental *Escherichia coli* Isolates. *Applied and Environmental Microbiology*, 77(22), p. 7915–7923.

Cid, D. et al., 2001. Association between intimin (eae) and EspB gene subtypes in attaching and effacing *Escherichia coli* strains isolated from diarrhoeic lambs and goat kids. *Microbiology,* Volume 147, pp. 2341-2353.

Clermont, O., Christenson, J. K., Denamur, E. & Gordon, D. M., 2013. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports,* 5(1), pp. 58-65.

Clermont, O., Gordon, D. & Denamur, E., 2015. Guide to the various phylogenetic classification schemes for *Escherichia coli* and the correspondence among schemes. *Microbiology,* 161(5), pp. 980-988.

Clermont, O. et al., 2011. Characterization of the cryptic *Escherichia* lineages: rapid identification and prevalence. *Environmental Microbiology,* 13(9), pp. 2468-2477.

Clermont, O. et al., 2008. Evidence for a human-specific *Escherichia coli* clone. *Environmental Microbiology,* 10(4), pp. 1000-1006.

Coli, D., Jospin, G. & Darling, A. E., 2015. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics (Oxford, England),* 31(4), pp. 587-589.

Coura, F. M. et al., 2015. Phylogenetic Group Determination of *Escherichia coli* Isolated from Animals Samples. *The Scientific World Journal,* Volume 2015.

Day, M. J. et al., 2016. Population structure of *Escherichia coli* causing bacteraemia in the UK and Ireland between 2001 and 2010. *Journal of Antimicrobial Chemotherapy,* 71(8), pp. 2139-2142.

de Muinck, E. J. et al., 2013. Comparisons of infant *Escherichia coli* isolates link genomic profiles with adaptation to the ecological niche. *BMC Genomics,* 14(81).

Dixit , O. V. A., O'Brien, C. L., Pavli, P. & Gordon, D., 2018. Within-host evolution versus immigration as adeterminant of *Escherichia coli* diversity in thehuman gastrointestinal tract. *Environmental Microbiology ,* 20(3), pp. 993-1001.

Donnenberg, M. S. et al., 1993a. Role of the eaeA gene in experimental enteropathogenic *Escherichia coli* infection. *The Journal of Clinical Investigation,* 92(3), pp. 1412-1417.

Donnenberg, M. S. et al., 1993b. The role of the eae gene of enterohemorrhagic *Escherichia coli* in intimate attachment in vitro and in a porcine model. *The Journal of Clinical Investigation,* 92(3), pp. 1418-1424.

Edberg, S. C., Rice, E. W., Karlin, R. J. & Allen, M. J., 2000. *Escherichia coli*: the best biological drinking water indicator for public health protection. *Journal of Applied Microbiology*, 88(s1), pp. 106S-116S.

EPA (Environmental Protection Agency) Report of Task Force on Guide Standard and Protocol for Testing Microbiological Water Purifiers. (1986).

Ferguson, C. M., Coote, B. G., Ashbolt, N. J. & Stevenson, I. M., 1996. Relationships between indicators pathogens and water quality in an estuarine system. *Water Research*, 30(9), pp. 2045-2054.

Field, K. G. & Samadpour, M., 2007. Fecal source tracking, the indicator paradigm, and managing water quality. *Water Research,* 41(16), pp. 3517-3538 .

Gerdes, K., 2016. Hypothesis: type I toxin–antitoxin genes enter the persistence field—a feedback mechanism explaining membrane homoeostasis. *Philosophical Transactions B*, 371(1707), p. 20160189.

Genoscope Microscope. Available at: http://www.genoscope.cns.fr/agc/microscope/search/blast.php? [Accessed 16 October 2017].

Gomi, R., Matsuda, T., Matsui, Y. & Yoneda, M., 2014. Fecal Source Tracking in Water by Next-Generation Sequencing Technologies Using Host-Specific *Escherichia coli* Genetic Markers. *Environmetal Science and Technology,* Volume 48, pp. 9616-9623.

Gordon, D. M. & Cowling, A., 2003. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology,* Volume 149, p. 3575–3586.

Gordon, D. M. et al., 2017. Fine-Scale Structure Analysis Shows Epidemic Patterns of Clonal Complex 95, a Cosmopolitan *Escherichia coli* Lineage Responsible for Extraintestinal Infection. *mSphere*, 2(3), pp. e00168-17.

Gordon, D. M., O'Brien, C. L. & Pavli, P., 2015. *Escherichia coli* diversity in the lower intestinal tract of humans. *Environmental Microbiology Reports,* 7(4), pp. 642-648.

Gordon, D. M., Stern, S. E. & Collignon, P. J., 2005. Influence of the age and sex of human hosts on the distribution of *Escherichia coli* ECOR groups and virulence traits. *Microbiology,* 151(1), pp. 15-23.

Grainge, I. & Sherratt, D. J., 1999. Xer Site-specific Recombination DNA strand rejoining by recombinase XerC. *Journal of Biological Chemistry,* 274(10), p. 6763–6769.

Greiner, R., Konietzny, U. & Jany, K. D., 1993. Purification and characterization of two phytases from *Escherichia coli*. *Archives of Biochemistry and Biophysics*, 303(1), pp. 107-113.

Hammer, O., Harper, D. A. T. & Ryan, P. D., 2001. PAST: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontologia Electronica*, 4(1), pp. 1-9.

Harwood, V. J. et al., 2014. Microbial source tracking markers for detection of fecal contamination in environmental waters: relationships between pathogens and human health outcomes. *FEMS Microbial Reviews,* Volume 38, pp. 1-40.

Hazen, T. H. et al., 2013. Refining the pathovar paradigm via phylogenomics of the attaching and effacing *Escherichia coli*. *Proceedings of the National Academy of Science,* 110(31), p. 12810–12815.

Heller, D. M., Tavag, M. & Hochschild, A., 2017. CbtA toxin of *Escherichia coli* inhibits cell division and cell elongation via direct and independent interactions with FtsZ and MreB. *PLOS Genetics,* 13(9), p. e1007007.

Ishii, S., Ksoll, W. B., Hicks, R. E. & Sadowsky, M. J., 2006. Presence and Growth of Naturalized *Escherichia coli* in Temperate Soils from Lake Superior Watersheds. *Applied and Environmental Microbiology,* 72(1), pp. 612-621.

Ishii, S. et al., 2014. Water quality monitoring and risk assessment by simultaneous multipathogen quantification. *Environmental Science and Technology,* 48(9), pp. 4744-4749.

Ishii, S. & Sadowsky, 2., 2008. *Escherichia coli* in the Environment: Implications for Water Quality and Human Health. *Microbes and Environments,* 23(2), pp. 101-108.

Jang, J. et al., 2017. Environmental *Escherichia coli*: ecology and public health implications-a review. *Journal of Applied Microbiology,* 123(3), pp. 570-581.

Johnson, J. R., 2002. Evolution of Pathogenic *Escherichia coli*. In: M. S. Donnenberg, ed. *Escherichia Coli-Virulence Mechanisms of a Versatile Pathogen.* Baltimore, Maryland: Elsevier BV, pp. 55-77.

Johnson, J. R., Delavari, P., Kuskowski, M. & Stell, A. L., 2001. Phylogenetic Distribution of Extraintestinal Virulence-Associated Traits in *Escherichia coli*. *The Journal of Infectious Diseases,* 183(1), pp. 78-88.

Johnson, J. R. & O'Bryan, T. T., 2004. Detection of the *Escherichia coli* Group 2 Polysaccharide Capsule Synthesis Gene kpsM by a Rapid and Specific PCR-Based Assay. *Journal of Clinical Microbiology,* 42(4), p. 1773–1776.

Johnson, J. R. & Russo, T. A., 2002. Extraintestinal pathogenic *Escherichia coli* : "The other bad E coli ". *Journal of Laboratory and Clinical Medicine* , 139(3), pp. 155-162.

Kaper, J. B., Nataro, J. P. & Mobley, H. L. T., 2004. Pathogenic *Escherichia coli*. *Nature Review Microbiology,* 2(2), pp. 123-140.

Katouli, M., 2010. Population structure of gut *Escherichia coli* and its role in development of extra-intestinal infections. *Iranian Journal of Microbiology,* 2(2), pp. 59-72.

Klemm, P., Tong, S., Nielsen, H. & Conway, T., 1996. The gntP Gene of *Escherichia coli* Involved in Gluconate Uptake. *Journal of Bacteriology,* 178(1), pp. 61-67.

Kobayashi, H. et al., 2003. Prevalence and Characteristics of eae-Positive *Escherichia coli* from Healthy Cattle in Japan. *Applied and Environmental Microbiology,* 69(9), pp. 5690-5692.

Land, M. et al., 2015. Insights from 20 years of bacterial genome sequencing. *Functional and Integrative Genomics,* 15(2), pp. 141-161.

Lange, R. & Hagge-Aronis, R., 1991. Identification of a central regulator of stationary-phase gene expression in *Escherichia coli*. *Molecular Microbiology,* 5(1), pp. 49-59.

Luo, C. et al., 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *PNAS,* 108(17), p. 7200–7205.

Markiewicz, L. H. et al., 2013. Diet shapes the ability of human intestinal microbiota to degrade phytate – in vitro studies. *Journal of Applied Microbiology,* 115(1), pp. 247-259.

Microbial Nucleotide Blast. Available at: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=MicrobialGenomes [Accessed 17 October 2017].

Mullaney, E. J., Daley, C. B. & Ullah, A. H., 2000. Advances in phytase research. *Advances in Applied Microbiology,* Volume 47, pp. 157-199.

NHMRC, N., 2011. *Australian Drinking Water Guidelines Paper 6 National Water Quality Management Strategy.* Canberra: National Health and Medical Research Council, National Resource Management Ministerial Council, Commonwealth of Australia.

Nowrouzian, F. L., Adlerberth, I. & Wold, A. E., 2006. Enhanced persistence in the colonic microbiota of *Escherichia coli* strains belonging to phylogenetic group B2: role of virulence factors and adherence to colonic cells. *Microbes and Infection,* 8(3), pp. 834-840.

Odonkor, S. T. & Ampofo, J. K., 2013. *Escherichia coli* as an indicator of bacteriological quality of water: an overview. *Microbiology research,* 4(1), p. e2.

Oh, S. et al., 2012. Genomic Diversity of Escherichia Isolates from Diverse Habitats. *PLoS ONE,* 7(10), p. e47005.

Page, A. J. et al., 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics,* 31(22), pp. 3691-3693.

Peleg, A. et al., 2005. Identification of an *Escherichia coli* Operon Required for Formation of the O-Antigen Capsule. *Journal of Bacteriology,* 187(15), p. 5259–5266.

Picard, B. et al., 1999. The Link between Phylogeny and Virulence in *Escherichia coli* Extraintestinal Infection. *Infection and Immunity,* 67(2), pp. 546-553.

Power, M. L. et al., 2005. Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. *Environmental Microbiology,* 7(5), pp. 631-640.

Quero, G. M., Fasolato, L., Vignaroli, C. & Luna, G. M., 2015. Understanding the association of *Escherichia coli* with diverse macroalgae in the lagoon of Venice. *Scientific Reports*, Volume 5, p. 5:10969.

Rodriguez , E., Han, Y. & Lei, X. G., 1999. Cloning, sequencing, and expression of an *Escherichia coli* acid phosphatase/phytase gene (appA2) isolated from pig colon. *Biochemical and Biophysical Research Communications,* 257(1), pp. 117-123.

Sathiyamoorthy, K., Mills, E., Franzmann, T. M. & Saper, M. A., 2011. The crystal structure of *Escherichia coli* group 4 capsule protein GfcC reveals a domain organization resembling that of Wza. *Biochemistry,* 50(24), pp. 5465-5476.

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics,* 30(14), pp. 2068-2069.

Serina, S. et al., 2004. Scanning the *Escherichia coli* chromosome by random transposon mutagenesis and multiple phenotypic screening. *Research in Microbiology,* 155(8), pp. 692-701.

Smati, M. et al., 2013. Real-Time PCR for Quantitative Analysis of Human Commensal *Escherichia coli* Populations Reveals a High Frequency of Subdominant Phylogroups. *American Society for Microbiology,* 79(16), pp. 5005-5012.

Subashchandrabose, S. et al., 2013. Genome-Wide Detection of Fitness Genes in Uropathogenic *Escherichia coli* during Systemic Infection. *PLOS Pathogens,* 9(12), p. e1003788.

Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E., 2010. The population genetics of commensal *Escherichia coli*. *Nature Reviews Microbiology,* Volume 8, pp. 207-217.

Thomassin, J. L. et al., 2013. Both Group 4 Capsule and Lipopolysaccharide O-Antigen Contribute to Enteropathogenic *Escherichia coli* Resistance to Human α-Defensin 5. *PLOS One,* 8(12), p. e82475.

Touati, E. & Danchin, A., 1987. The structure of the promoter and amino terminal region of the pH 2.5 acid phosphatase structural gene (appA) of *E. coli*: a negative control of transcription mediated by cyclic AMP. *Biochimie,* 69(3), pp. 215-221.

Touchon, M. et al., 2009. Organised genome dynamics in the *Escherichia coli* species results in high diverse adaptive paths. *PLOS Genetics,* 5(1), p. 5:e1000344.

van Elsas, J. D., Semenov, A. V., Costa, R. & Trevors, J. T., 2011. Survival of *Escherichia coli* in the environment: fundamental and public health aspects. *The ISME Journal,* Volume 5, pp. 173-183.

Vignaroli, C. et al., 2015. Adhesion of marine cryptic Escherichia isolates to human intestinal epithelial cells. *International Society for Microbial Ecology,* Volume 9, pp. 508-515.

Wade, T. J. et al., 2008. High sensitivity of children to swimming-associated gastrointestinal illness: results using a rapid assay of recreational water quality.. *Epidemiology,* 19(3), p. 375–383.

Walk, S. T. et al., 2007. Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. *Environmental Microbiology,* 9(9), pp. 2274-2288.

Walk, S. T. et al., 2009. Cryptic lineages of the genus Escherichia. *Applied and Environmental Microbiology,* 75(20), p. 6534–6544.

Ahmed, W. et al., 2015. Assessment of Genetic Markers for Tracking the Sources of Human Wastewater Associated *Escherichia coli* in Environmental Waters. *Environmental Science and Technology,* 49(15), pp. 9341-9346.

Whitfield, C., 2006. Biosynthesis and Assembly of Capsular Polysaccharides in *Escherichia coli*. *Annual Review of Biochemistry,* Volume 75, pp. 39-68.

Whitfield, C. & Roberts, I., 2002. Structure, assembly and regulation of expression of capsules in *Escherichia coli*. *Molecular Microbiology*, 31(5), pp. 1307-1319.

Yoon, S. M. et al., 2011. Transgenic microalgae expressing *Escherichia coli* AppA phytase as feed additive to reduce phytate excretion in the manure of young broiler chicks. *Applied Microbiology and Biotechnology*, 91(3), pp. 553-563.

CHAPTER 4

Survival of *E. coli* phylogroup B2 isolates in water

Introduction

Escherichia coli is known to have a biphasic lifestyle consisting of host-associated and host-independent phases (Van Elsas et al., 2011). Savageau (1983), considered the primary habitat of *E. coli* to be the gastrointestinal tracts of mammals, while water, soil and sediment are considered as the species secondary habitat. In its primary habitat *E. coli* achieves cell densities of 10^5 to 10^7 colony-forming units (CFU) per gram of faeces, and an isolate may persist in a host for a few days or many years. Previous studies suggest that *E. coli* typically has a life span of 1-5 days in external environments such as water and soil (Hartl and Dykhuizen, 1984).

Traditionally, *E. coli* is considered a faecal indicator bacteria (FIB) and used to detect the presence of recent faecal contamination indicating that other pathogens may also be present (Wolf, 1972; WHO, 1993; Leclerc et al., 2001). It is used as FIB mainly due to the belief that it is commonly present in high numbers in the G-I tract of mammals, and when in the secondary environment it has a short life span with little or no cell division (Boehm et al., 2009; Harwood et al., 2014). *E. coli* is also easily culturable under laboratory conditions. Some studies, however, have suggested that not only do some strains have the ability to persist for long periods in water (a low nutrient and stressful environment) even in the absence of direct faecal contamination, but they may also undergo significant growth, achieving cell densities in excess of 10^4 CFU/ 100ml (Byrd and Colwell 1993; Vital et al., 2008; Ihssen and Egli, 2005; Gordon, 2001; Power et al., 2005). These findings challenge the use of *E. coli* as an indicator organism (van Elsas et al., 2011).

For the century or more that *E. coli* has been used as FIB, it has not been known that *E. coli* can enter a dormant state known as Viable but Not-Culturable (VBNC) in order to survive the stresses present in external environments. However, Xu et al. (1982) was the first to report that *E. coli* can enter the VBNC state. Various studies over the past three decades have also shown *E. coli* to enter into a VBNC state (Barcina et al., 1990; Arana

et al., 2007; Asakura et al., 2007; Na et al., 2006; Juhna et al., 2007b; Liu et al. 2008; Zhao et al., 2013).

The VBNC state is a state where cells are alive, but dormant and unable to form colony-forming units under adverse or stress conditions (McKay, 1992; Colwell, 2009; Pienaar et al., 2016). The VBNC state may be activated when cells are exposed to various stress conditions such as temperatures below the optimum range for growth, elevated osmotic concentration, starvation, toxic free radicals and metals such as cadmium, copper, lead, mercury or pH changes (Klein and Alexander, 1986; Oliver, 2000). The VBNC state is distinct from both normal live culturable cells and dead cells, both metabolically and physiologically. Compared with dead cells, VBNC cells have intact membranes, and successful gene expression with no cytoplasmic leakage. In dead cells, there is no gene expression due to ruptured cell membranes (Xu et al., 1982; Wang and Doyle, 1997; Li et al., 2014). Compared to cells that are culturable and detected on suitable laboratory media, VBNC cells are not culturable under routine conditions (Oliver, 2000). Typically, VBNC cells also have lower metabolic activity and differ in outer membrane composition compared to live culturable cells (Muela et al., 2004; Li et al., 2014).

Studies also indicate that after a variable length of time, *E. coli* in a VBNC state can become culturable either spontaneously or following induction. Induced resuscitation is done by adding growth supplements or removing the stress that triggered the initial VBNC response (Ding et al.,2016; Zhang et al., 2015; Reissbrodt et al., 2002). In some VBNC *E. coli* regrowth is stimulated by signal molecules from adjoining cells or their by-products (Cuny et al., 2005). Other studies have confirmed that some VBNC cells are unable to recover after initial loss of culturability and die (Pinto et al., 2011; Arana et al., 2007). Cells that die after being in the VBNC state cannot induce disease (Sachidanandham and Gin, 2009). But, the cells of some pathogenic strains in the VBNC state maintain their virulence potential after their recovery from being VBNC (Makino et al., 2000; Zhao et al., 2016).

In water, when the *E. coli* is in VBNC state, it cannot be used as FIB (Oliver et al., 2005; Pienaar et al., 2016; Abberton et al., 2016) as it is difficult/impossible to grow cells in the standard laboratory media used by the water industry (Abberton et al., 2016; Ward et al., 1990; Blackburn and McCarthy, 2000; Keer and Birch, 2003; Juhna et al., 2007a; van Elsas et al., 2011). Failure to detect *E. coli* cells in the VBNC state may lead to false

negatives, which in turn result in a greater risk to public health (Liu et al., 2010; Li et al., 2014; Anderson et al., 2004; Cappelier et al., 2007).

Although *E. coli* is considered a well-studied model organism, its survival in the external environment is poorly understood. A previous study suggested that cell survival in external environments varied with the phylogenetic membership of the isolate, with phylogroup A and B1 isolates surviving better than B2 or D isolates in aquatic environments (Berthe et al., 2013). In addition, B2 strains are the least likely of the major *E. coli* phylogroups to be detected in water (Power et al., 2005; Castro Stoppe et al., 2017). By contrast, they are the phylogroup that is most likely to be recovered from faecal samples especially from humans living in developed countries (Picard et al., 1999; Gordon and Cowling, 2003; Gordon et al., 2005; Nowrouzian et al., 2006; Escobar-Paramo et al., 2006; Le Gall et al., 2007; Clermont et al., 2013; Smati et al., 2015; Gordon et al., 2015).

The aim of the present study was to examine the survival of *E. coli* strains belonging to phylogroup B2 isolated from humans, birds, mammals, and water samples and to determine if variation in among-strain survival might be explained by variation in the variable gene content of the isolates.

Materials and Methods

Isolate Selection

Fifty *E. coli* isolates belonging to phylogroup B2 were chosen for the water survival experiment. These isolates represented 29 Sequence types (ST) and were isolated from humans, birds, mammals, and water samples (Gordon and Cowling, 2003; Power et al., 2005; Blyton et al., 2015; Gordon et al., 2015; Blyton and Gordon, 2017; Dixit et al., 2018) (Table 4.1).

Table 4.1 List of *E. coli* phylogroup B2 strains used in this study, their corresponding source of isolation, sequence type and serotype.

NAME	SOURCE	ST	Serotype
B004	BIRD	91	O39:H4
B103	BIRD	1894	O13:H5
B108	BIRD	73	O50/O2:H1
B127	BIRD	131	O25:H4
B1547	BIRD	131	O25:H4
B288	BIRD	127	O6:H31
B339	BIRD	978	O83:H27
B377	BIRD	1899	O4:H40
B620	BIRD	2622	O83:H6
20-5-R7	HUMAN	73	O50: H1
47_1_TC4	HUMAN	110	O99: H4
57_5_R8	HUMAN	537	O75:H5
58-2-HC1	HUMAN	28	O96: H7
60-1T11	HUMAN	80	O75: H7
62-1TI3	HUMAN	12	O4:H5
69-1-TI1	HUMAN	569	O134: H31
H001	HUMAN	681	O8:H10
H112	HUMAN	95	O1:H7
H223	HUMAN	141	O50/O2: H6
H437	HUMAN	95	O50/O2:H7

NAME	SOURCE	ST	Serotype
H504	HUMAN	95	O18:H7
H522	HUMAN	3276	O131:H6
H578	HUMAN	1257	O8:H10
M0528	MAMMAL	1858	O75:H5
M0549	MAMMAL	429	O83:H4
M605	MAMMAL	1876	O39:H4
POSS-24	MAMMAL	141	O50/O2: H6
POSS-70	MAMMAL	3307	O170:H4
TA098	MAMMAL	1257	O8:H10
TA206	MAMMAL	1386	O13/O135:H4
TA258	MAMMAL	3276	O131:H6
TA265	MAMMAL	80	O7:H7
TA309	MAMMAL	681	O8:H10
E2059	WATER	95	:H7
E2062	WATER	3291	O25:H5
E2549	WATER	1858	O6:H5
E3317	WATER	28	O177:H6
E4259	WATER	636	O83:H7
E4453	WATER	135	O83:H1
E4931	WATER	3307	O170:H5
E5598	WATER	1899	O4:H40
E6649	WATER	1386	O13:H4
E7087	WATER	95	O18:H7
E7242	WATER	681	O8:H10
E7253	WATER	3646	O16:H14
E7603	WATER	569	O134:H31
E7615	WATER	95	O50/O2:H7
E7727	WATER	3307	O170:H5
E8621	WATER	28	:H6
E9644	WATER	1873	:H4

Water Treatments for Microcosms

Fresh water from an artificial pond at the Research School of Biology, Australian National University, was used for the survival experiments. The pond was about 650 cm x 870 cm and 70 cm deep and was not connected to a filtration system. The pond was home to a variety of emergent and submergent vegetation, fishes, turtles and water dragons. Water was collected in sterilised bottles and allowed to stand for 24 hours to enable large particulate matter to settle. The samples were then gravity filtered 3 times through 15.0 cm diameter, 11 μm pore size, Grade 1 cellulose Whatman® filter paper. The filter papers were folded into a fluted shape for effective filtration. This filtered raw pond water was then split into two batches. Half of the water was heat sterilised by autoclaving at 121°C for 20 mins. The other half was filter sterilised using aPES membrane-filtering unit (Catalogue No.: 566-0021) from Thermo Scientific™ Nalgene™ Rapid-Flow™ having a pore size of 0.22 μm. The pH of the filtered sterilise and autoclaved pond water was 7.

Microcosms

Working inoculums were prepared by removing an aliquot of stock cultures that were stored in glycerol (50% v/v) at -80°C and dilution streaking onto a MacConkey agar plate that was then incubated overnight at 37°C. Cells from a single colony of each isolate were inoculated into individual flasks containing 5ml Lysogeny Broth that were incubated at 37°C, overnight with shaking at 170 rpm to obtain a cell density of $\approx 10^9$ CFU/ml. After overnight incubation, the culture was serially diluted in 0.9% saline to obtain a cell density of 10^5 CFU/ml. A 100μl aliquot from the 10^5CFU/ml dilution was added to test tubes containing one of 9.9 ml of autoclave sterilised pond water (APW), filter sterilised pond water (FPW), and autoclaved Deionized Water (DW) (Fig 4.1). The tubes were mixed thoroughly by vortexing to obtain an even distribution of bacterial population, loosely capped, and incubated at 20°C without shaking. Prior to each sample being removed the tubes were briefly vortexed for 10 secs. The tubes were sampled by spread plating a 100μl aliquot of each microcosm onto Luria Bertani (LB) Agar plates that were then incubated overnight at 37°C, and the number of colonies were counted using ProtoCOL 3 colony counter. The plating was done daily until the number of colony forming units reached zero at which point the microcosms were sampled 3x per week.

The APW and DW microcosms were sampled for 115 days, while the FPW microcosms were sampled for 60 days.

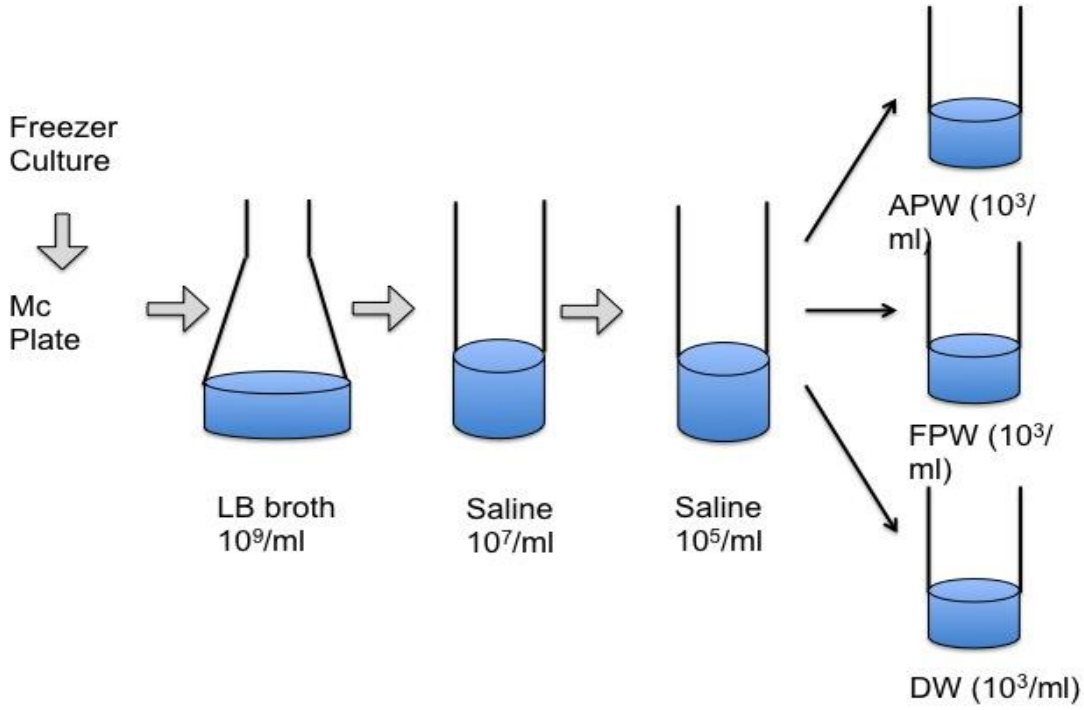


Fig 4.1. Microcosm experimental setup

Repetitive Element Palindromic (REP) PCR

At the end of the sampling period for each APW microcosm in which viable cells were still detected, surviving cells were fingerprinted and compared with the fingerprint of the isolate used to initiate the microcosm. DNA extraction was performed using DNAzol (Molecular Research Center Inc.) and a 200µl aliquot of an overnight Lysogeny Broth culture following the manufacturer's protocol. REP typing was done using Enterobacterial Repetitive Intergenic Consensus (ERIC) PCR with 10µM of the ERIC primer (Versalovic et al., 1991), and 2.5 U Platinum Taq (Invitrogen) prepared in 5x buffer (BIOLINE) with 1.2µl of template DNA to a final volume of 20µl. The PCR reaction included an initial denaturation step at 95°C for 2 minutes, followed by 30 cycles of denaturation at 94°C for 3 seconds and 92°C for 30 seconds, annealing at 50°C for 1 minute, extension at 65°C for 8 minutes and a final extension at 72°C for 8 minutes. The PCR product was run on 1.2% agarose gel in TBE buffer. The amplified DNA fragments

in the gel were visualised and captured using BIO-RAD GelDOC™ imaging system with UV transilluminator.

Ammonia (NH_3/NH_4^+), Phosphate (PO_4^{3-}) and Nitrate (NO_3^-) Test

The water used for these experiments was tested to determine the concentration of ammonia, phosphate and nitrate in each sample: raw pond water (RPW) - the freshwater that has been paper filtered 3x using Whatman® filter paper, filtered pond water (FPW)- freshwater that was further filtered using a 0.22µm filter, autoclaved pond water (APW)- freshwater that was heat sterilised, and autoclaved deionised water (DW). Analyses were carried out using colorimetric kit method from API® (USA): ammonia test kit (Catalogue No.: APH151), nitrate test kit (Catalogue No.: APH150), phosphate test kit (Catalogue No.: APH229) according to the manufacturer's protocols.

Mass Spectrometry

Extraction of water metabolites using mass spectrometry was done by aliquoting 5ml of APW and FPW into separate test tubes. 500µl of 200µg/L rabitol was added to each test tube and vortexed vigorously for a minute. 2ml of the water-rabitol mix was transferred into a separate test tube and supplemented with 2ml of 100% methanol followed by vigorously vortexing for 1 minute. The tubes were then centrifuged at 180 rpm for 20 minutes. A 1 ml aliquot of the supernatant was added to 1.5 ml a microfuge tube and allowed to dry completely for approximately 3 to 5 hours at 40°C in a vacuum drying chamber. After drying, 200µl of methanol was added to the dried tubes and the mixture was vortexed completely, then transferred to GC vials and dried again completely. Derivatizing agents such as 10µl of Methoxylamine (Sigma-Aldrich 226904-5G) and 15µl of MSTFA (N-Methyl-N-(trimethylsilyl)-trifluoroacetamide) (Sigma-Aldrich 394866-10X) were added to each of the GC vials to determine the compounds present. The samples were then analysed using TRACE™ Gas Chromatography and ThermoPolarisQ™ Mass Spectrometry (GC-MS). The results of each compound detected were analysed using ANALYSER PRO with 70% confidence (Draper et al., 2004).

Statistical analysis

Survival and cell recovery patterns are analysed using R Studio® version 2.7.0 (https://www.R-project.org). The R packages installed were knitr and markdown and the libraries used were survival, ggplot2, survminer, lme4, lmerTest, lsmeans.

Whole genome sequence data generated using Illumina Nextera library preparation kits and sequenced with the Illumina MiSeq platform using V3 chemistry was available for each of the strains. The sequence reads were assembled using A5MiSeq (Coli et al., 2015), annotated using Prokka (Seemann, 2014) and a pangenome analysis was conducted using Roary (Page et al., 2015). The size of the pan genome was determined to be 15173 genes; of these, only those with a frequency of >20% and <80% were selected to study potentially informative genes of significance as > 80 % and < 20 % are the most common and most rare genes that may cause statistical insignificance to the data. Hence, this resulted in 1288 genes to be used in subsequent analyses.

Additional analyses were carried out to determine if differences in variable gene content of the strains explained any of the variation in (i) the rate at which strains loss culturability (ii) the time taken for strains to recover culturability in APW and (iii) strains' death rate in FPW. The first step in these analyses was to do a Principal Coordinate Analysis (PCoA) based on the variable gene content of the strain, then to ask if any of the PCO axes explained any of the variation in the attribute being investigated. If no variation could be explained then no further analyses were pursued. If some of the variation was explained then Partial Least Square (PLS) analysis was used determine a variables importance score (VIP). The three most important variables were selected and used in Analysis of Variance models. The PCoA analysis, PLS analysis, Distribution analysis, Fit model were all computed using the JMP® software, version 13.2.0 (Jones and Sall, 2011). The genes were further annotated using Microbial Blast Search (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=MicrobialGenomes).

Results

The survival experiment was started with the working assumption, suggested by the literature, that phylogroup B2 strains have a short life span in fresh water habitats when

compared to their primary habitat. In this study, 29 different STs of phylogroup B2 from sources such as humans, birds, mammals, and water were used. A few of these isolates were from ExPEC lineages such as ST73 (2), 95 (6), 131(2).

Water Chemistry

Ammonia (NH_3/NH_4^+), phosphate (PO_4^{3-}) and nitrate (NO_3^-) analyses indicated that nitrate and phosphate concentrations were below the limits of detection for all water samples. Ammonia levels were below the detection limit for all water samples, with the exception of the APW samples, where the ammonia concentration was 1.00 ppm (mg/L).

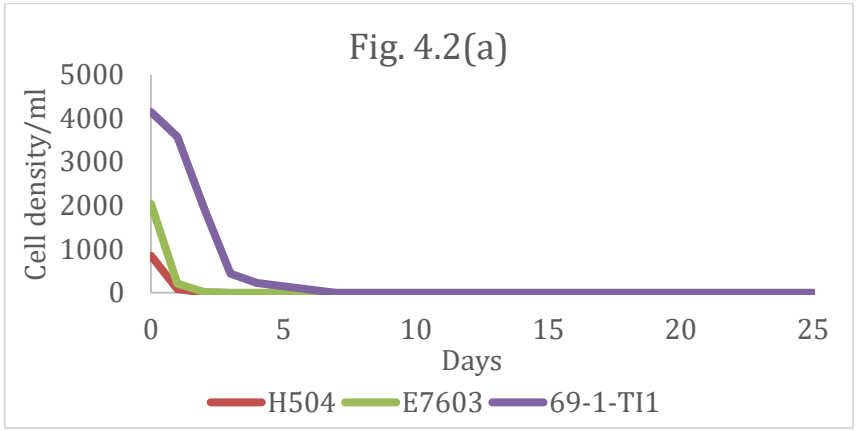
Mass spectrometry analysis of the composition of APW and FPW revealed that three different compounds were found to be uniquely present in APW and not in FPW. Similarly, five other compounds were exclusive to FPW and not in APW (Table 4.2)

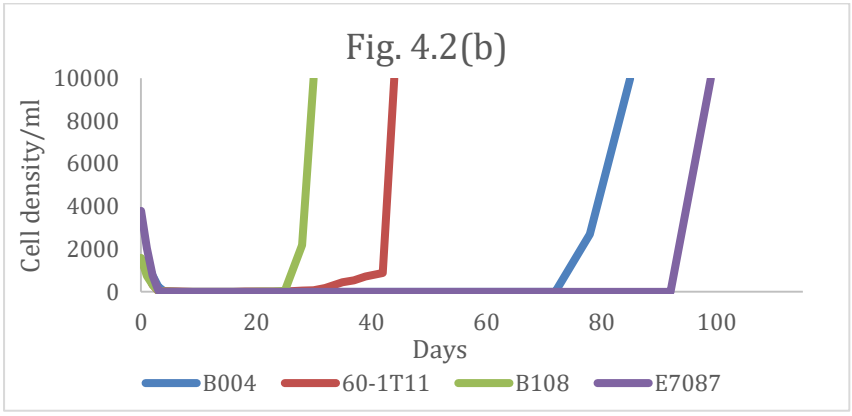
Table 4.2 Difference in compounds between APW and FPW using mass spectrometry analysis

Compounds only in APW	Compounds only in FPW
1. Butanoic acid, 2,4 bid[(trimethylsilyl)oxy]-trimethylsilyl ester	1. 5-Trimethylsilyloxy-n-valeric acid, trimethylsilyl ester
2. D-Pinitol, pentakis (trimethylsilyl)ether	2. D-Xylopyranose, 1,2,3,4-tetrakis-O-(trimethylsilyl)
3. 2-Methylacetoacetic acid, di(trimethylsilyl) deriv.	3. Galactose oxime hexakis (trimethylsilyl)
	4. Pentitol, 1-desoxytetrakis-O-(trimethylsilyl)
	5. Sulfurous acid, 2-ethylexyl hexyl ester

Cell Survival in Autoclaved Pond Water

Monitoring of the number of CFUs/ml in the microcosms revealed three basic patterns (Fig. 4.2). They were (i) cells rapidly lose culturability and viable colony forming units were never again observed during the 115 days of sampling (Fig. 4.2a) (ii) cells rapidly lose culturability and the number of CFUs/ml declined to below the limit of detection, but after varying periods colony forming units were again observed (Fig. 4.2b) (iii) colony forming units were observed throughout the course of the experiment (Fig. 4.2c).





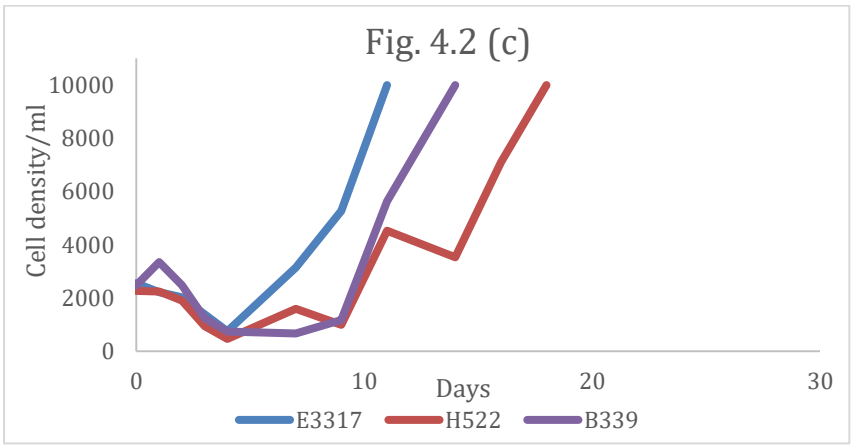


Fig. 4.2 Change in the number of CFUs/ml for phylogroup B2 strains in the autoclaved pond water microcosms. The data is presented for a representative subset of the strains. (a) viable cell counts declined to 0 and viable cells were never observed again (b) viable cell counts declined to 0 and eventually viable cells were subsequently again observed (c) viable cells were always observed.

For 29 of the 50 B2 isolates the number of colony forming units steadily declined until no colony forming units could be detected, but after a varying number of days colony

forming units were again observed, and eventually the number of colony forming units per ml exceeded the number of cells used to initiate the microcosm. For these 29 isolates, it took an average of 5.9 days to lose culturability. The average number of days between the loss of culturability and the return to culturability was 28 days, but ranged from 1 – 95 days (Fig. 4.3; Table 4.3).

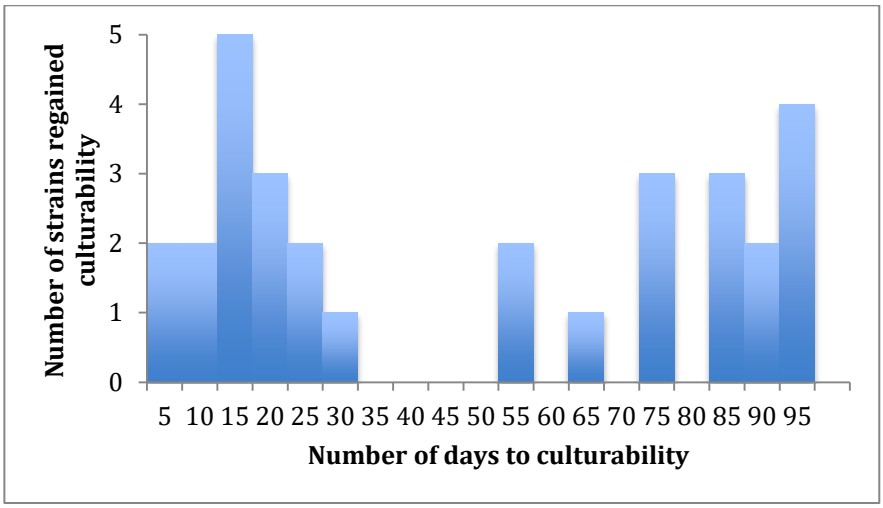


Fig. 4.3 Distribution of the number of days it took cells to return to being able to form colony forming units in the autoclaved pond water microcosms.

Table 4.3 The average number of days taken for the number of colony forming units per ml to decline to 0 ('lifespan') for phylogroup B2 strains in autoclaved pond water microcosms and the number of days it took each strain to regain the ability to form colony forming units.

Sample ID	SOURCE	ST	'Life Span' days	Recovery Time* days
B620	BIRD	2622	1.2	4
B108	BIRD	73	1.3	14
B1547	BIRD	131	1.4	62
B004	BIRD	91	1	71
B288	BIRD	127	1.4	74
B127	BIRD	131	1.6	81
B339	BIRD	978	0.2	No cD=0*
B103	BIRD	1894	0	No cR*
B377	BIRD	1899	1.8	No cR
57_5_R8	HUMAN	537	0.3	7
60-1T11	HUMAN	80	0.7	9
62-1TI3	HUMAN	12	1.6	9
H001	HUMAN	681	0.7	18

Sample ID	SOURCE	ST	'Life Span' days	Recovery Time* days
H437	HUMAN	95	1.4	55
58-2-HC1	HUMAN	28	1	56
H578	HUMAN	1257	0.8	76
20-5-R7	HUMAN	73	0	84
H112	HUMAN	95	0.1	95
H522	HUMAN	3276	0.4	No cD=0
H223	HUMAN	141	0.6	No cD=0
H504	HUMAN	95	2.2	No cR
69-1-TI1	HUMAN	569	0.7	No cR
47_1_TC4	HUMAN	110	0.9	No cR
TA265	MAMMAL	80	1.3	1
POSS-24	MAMMAL	141	1.2	2
M0528	MAMMAL	1858	1.4	2
M605	MAMMAL	1876	0.8	3
TA309	MAMMAL	681	0.9	12
TA258	MAMMAL	3276	1.3	74
M0549	MAMMAL	429	1.2	92
TA098	MAMMAL	1257	1.3	95
POSS-70	MAMMAL	3307	1.2	No cR
TA206	MAMMAL	1386	0	No cR
E4259	WATER	636	0.6	2
E7242	WATER	681	0.8	2
E5598	WATER	1899	1.7	3
E2059	WATER	95	1.1	7
E7615	WATER	95	0.4	85
E7087	WATER	95	1.3	95
E3317	WATER	28	0.3	No cD=0
E7603	WATER	569	2.3	No cR
E2062	WATER	3291	0	No cR
E2549	WATER	1858	0.5	No cR
E7727	WATER	3307	1.6	No cR
E4931	WATER	3307	0.5	No cR
E6649	WATER	1386	1.1	No cR
E7253	WATER	3646	0	No cR
E4453	WATER	135	2.2	No cR
E8621	WATER	28	0	No cR
E9644	WATER	1873	2.7	No cR

*Recovery Time- difference in days between isolates reaching zero cell density and recovery to colony forming units; No cR: the number of colony forming units declined to zero, but cells never regained culturability over 115 days of sampling. No cD=0: although initially declined, the number of colony forming units rapidly increased and was always >0.

Rep-PCR (ERIC) was performed on isolates that recovered after loss of cuturability to confirm that the cells present at the end of the experiment had the same genotype as the cells used to initiate the experiment. All the isolates had the same REP finger print except one (isolate H578 & H578R) and this isolate was removed from subsequent analyses (Fig. 4.4).

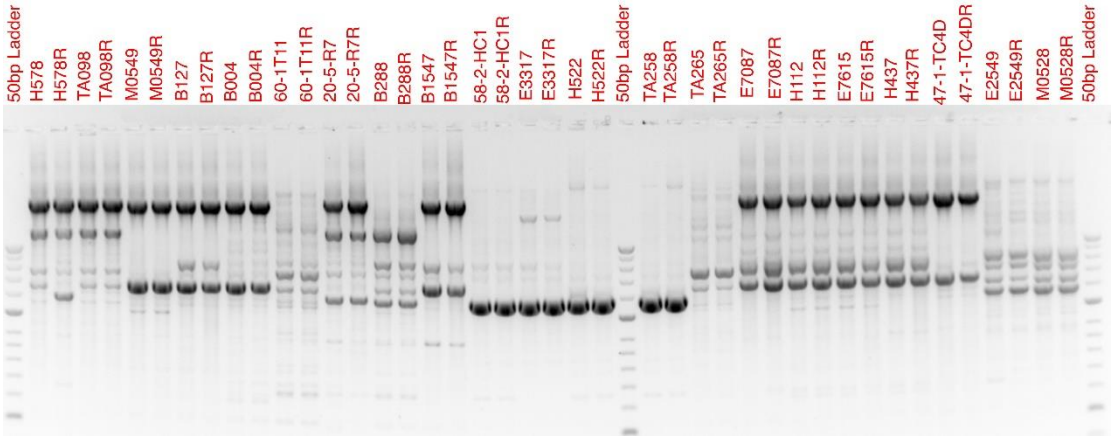


Fig. 4.4 rep-PCR DNA fingerprint pattern of *E. coli* isolates from initial inoculation at day zero and DNA after recovery to culturability (represented with R at the end of isolate ID)

For 17 of the 50 isolates, the number of colony forming units steadily declined until no colony forming units could be detected and colony forming units were never again detected over the balance of the 115 days of sampling. They also did not recover after the addition of 1% LB broth at the end of the experimental period. For these strains, it took an average of 4.1 days for these strains to lose their culturability. Four of the 50 isolates never reached a CFU/ml of 0 over the 115 days of the experiment. Strains that never regained culturability lost culturability significantly more rapidly than those eventually regained culturability, 4.1 vs 5.9 days respectively (Kruskal-Wallis Test; Prob>|Z|=0.028).

The fraction of strains failing to recover from their loss of culturability varied with the source of the isolates (Table 4.4) (Contigency Table Analysis: $p>X^2 = 0.0097$). Overall, 59% of the 17 water isolates never recovered the ability to form colony forming units, while, on average, only 22% of the isolates from a vertebrate host did not regain culturability. There was no effect of isolate source on the number of days to regain culturability (Kruskal-Wallis Test; Prob>|X^2|=0.54).

Table 4.4 Proportion of isolates that recovered or failed to recover from losing culturability in autoclaved pond water with respect to the source of the isolate.

SOURCE	No. of isolates	Isolates recovered from cD=0 (%)	Isolates with no recovery from cD=0 (%)	Average time to recovery (days)
Human	13	61.6%	23.1%	84
Bird	9	66.7%	22.2%	86
Mammal	10	80.0%	20.0%	77
Water	17	35.3%	58.8%	90

Deionised Water

In the deionised water microcosms, the cells density declined steadily irrespective of the source of isolates. On an average, the loss of culturability took 12.5 days and colony forming units were never subsequently observed. (Fig. 4.5; Table 4.5).

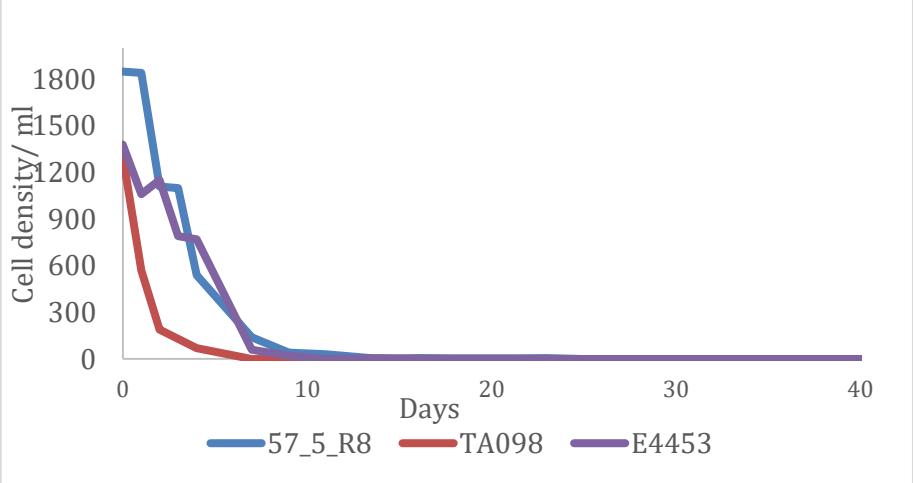
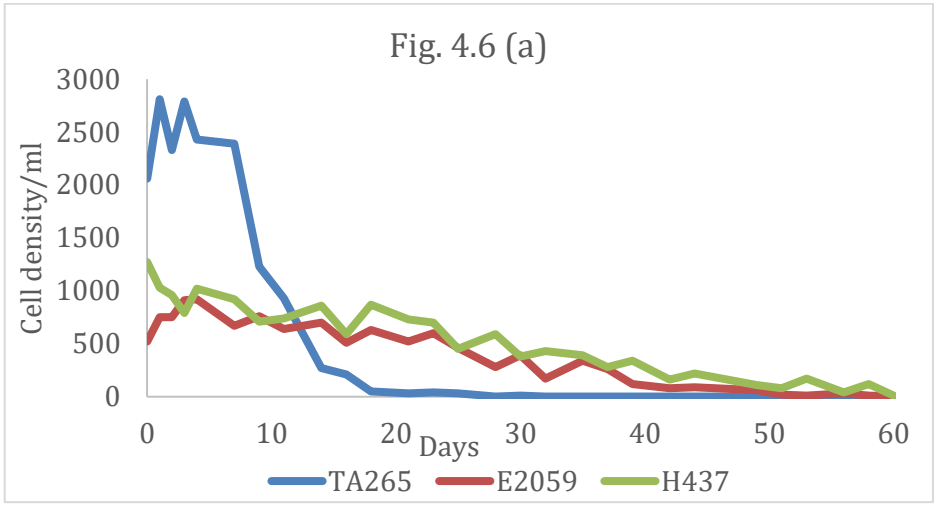


Fig. 4.5 Change in the number of CFUs ml^{-1} for a subset of phylogroup B2 strains in the deionized water microcosms. Only the results for the first 40 days of sampling are presented although the microcosms were sampled for 60 days.

The rate at which cells lost viability differed between DW and isolates that did not recover in APW. Interestingly, isolates in APW microcosms lost viability quicker (4.1 days) than isolates in DW microcosms (12.5 days) (Kruskal-Wallis Test; Prob>|Z|= <0.0001).

Filtered Pond Water

The survival of isolates in the FPW microcosms was very different to that observed either in the DW or APW microcosms, and exhibited two different patterns: no change in the number of CFUs/ml could be detected, or the number of CFUs/ml declined slowly (Fig. 4.6). Over the 60 days of sampling, the number of CFUs/ml declined to less than the limit of detection for only one isolate from a mammal. At the end of experiment, no colony forming units for this strain were detected after the addition of 1% Lysogeny Broth, indicating that the loss of culturability represented cell death rather than the cells entering a VBNC state.



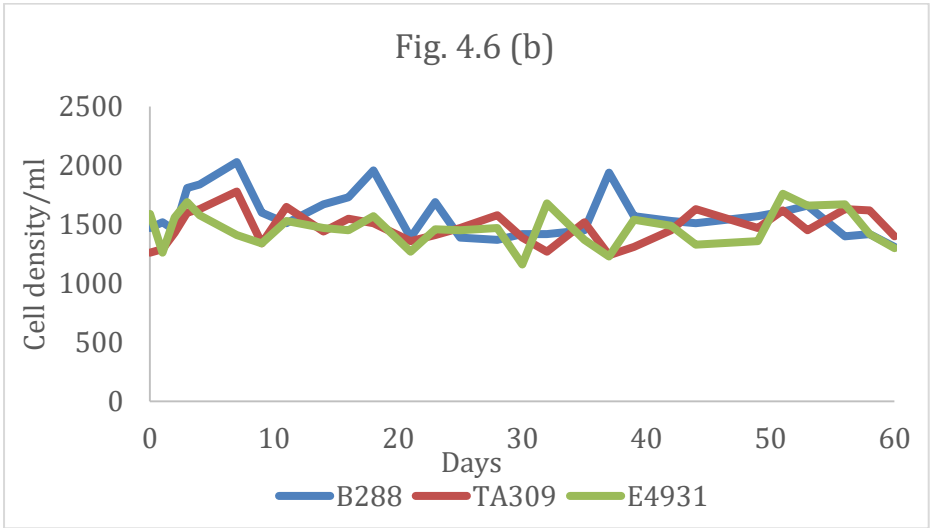


Fig. 4.6 Changes in the number of CFUs ml^{-1} over time for a subset of isolates in the filtered pond water microcosms. (a) isolates with very slow death rate (b) isolates with virtually constant cell densities.

The lifespan of cells in the FPW microcosms was highly variable and ranged from < 5 days to > 1 year. The source of the isolate explained none of the variation in the lifespan of the isolates (Kruskal-Wallis Test; Prob>X^2=0.78)

Table 4.5 Estimated number of days required for cells to lose culturability ('lifespan') in the deionised water (DW) and in filtered pond water (FPW) microcosms.

Sample ID	SOURCE	ST	DW 'Lifespan' days	FPW 'Lifespan' days
B127	BIRD	131	3.4	223
B004	BIRD	91	1	38
B108	BIRD	73	1.2	48
B288	BIRD	127	1.4	469
B1547	BIRD	131	1.3	186
B620	BIRD	2622	1.4	1647
B377	BIRD	1899	1.5	286
B103	BIRD	1894	2.9	151
B339	BIRD	978	2.7	2326
H578	HUMAN	1257	1.1	506
60-1T11	HUMAN	80	1.2	363
20-5-R7	HUMAN	73	1.1	47
58-2-HC1	HUMAN	28	0	135
H001	HUMAN	681	1.1	330
H112	HUMAN	95	2.1	529
H437	HUMAN	95	1.4	19
62-1TI3	HUMAN	12	2.8	216
57_5_R8	HUMAN	537	3.2	935
H504	HUMAN	95	3.2	71
69-1-TI1	HUMAN	569	1.1	197
47_1_TC4	HUMAN	110	31.3	138
H522	HUMAN	3276	2.2	151
H223	HUMAN	141	1.4	51
TA098	MAMMAL	1257	1.3	78

Sample ID	SOURCE	ST	DW 'Lifespan' days	FPW 'Lifespan' days
M0549	MAMMAL	429	1.0	135
M605	MAMMAL	1876	1.0	33
TA258	MAMMAL	3276	2.0	246
TA265	MAMMAL	80	3.1	5
TA309	MAMMAL	681	1.7	2386
POSS-24	MAMMAL	141	1.1	373
M0528	MAMMAL	1858	3.3	184
POSS-70	MAMMAL	3307	2.7	147
TA206	MAMMAL	1386	2.7	497
E4259	WATER	636	1.8	503
E7242	WATER	681	1.3	132
E7087	WATER	95	2.3	28
E2059	WATER	95	2.9	14
E7615	WATER	95	2.7	123
E5598	WATER	1899	0.9	550
E7603	WATER	569	1.6	207
E2062	WATER	3291	2.5	397
E2549	WATER	1858	2.5	69
E7727	WATER	3307	3.6	219
E4931	WATER	3307	3.7	8710
E6649	WATER	1386	1.8	178
E7253	WATER	3646	1.9	178
E4453	WATER	135	2.6	576
E8621	WATER	28	3.3	169
E9644	WATER	1873	2.1	55
E3317	WATER	28	3.8	384

Genetic exploration studies

There were 1288 variable genes with a frequency of between 20% and 80% among the 50 B2 strains. The dimensions of the presence/absence gene content matrix were reduced using a principal component analyses and the first 6 axes of the PCO were saved. None of the PCO axes were found to explain among isolate variation in the rate at which isolates lost culturablity in the APW experiment (Table 4.6). Hence no further analyses were done for the loss of culturability in APW.

Table 4.6 First six PCO scores based on the variation of cell culturability observed in APW and FPW

PCO	%Variation explained	Death rate in APW, P>F	Time to recovery in APW, P>F	Death rate in FPW, P>F
1	15.193	0.258	0.587	0.013
2	13.101	0.181	0.062	0.103
3	10.228	0.389	0.078	0.478
4	8.291	0.531	0.013	0.345
5	8.254	0.204	0.173	0.309
6	5.989	0.154	0.212	0.011

However, PCO axis 4 was found to explain a significant amount the variation in the number of days it took an isolate to return to culturability after it has lost the ability to produce colony forming units, while PCO axes 2 and 3 were borderline significant (Fig. 4.7).

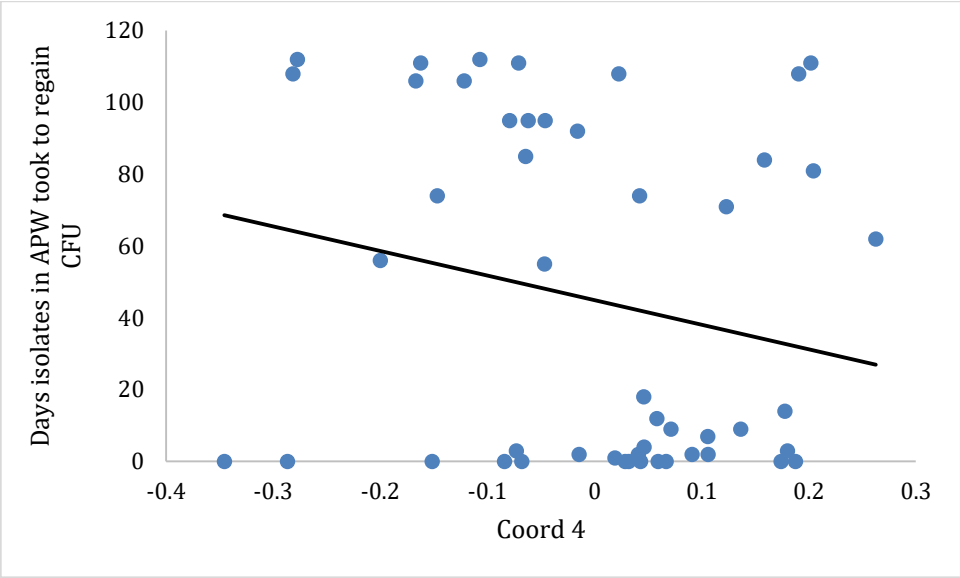


Fig. 4. 7 PCO score correlation between principle coordinate axis 4 and days to recovery from zero CFU of isolates in APW

Similary, PCO axis 1 was found to explain a significant amount the variation in the rate at which isolates lost their ability to produce colony forming units in FPW (Fig. 4.8).

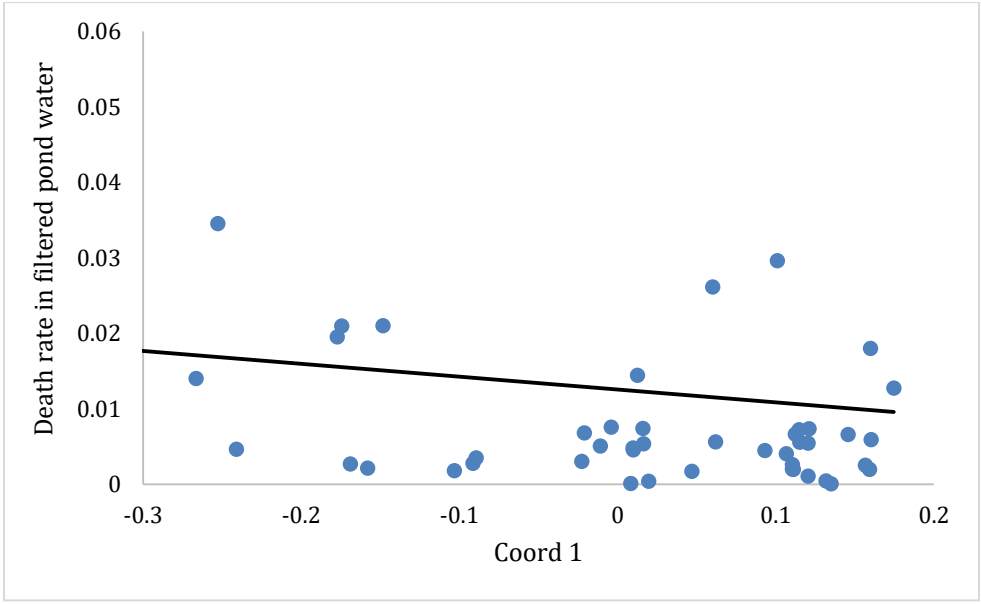


Fig. 4. 8 PCO score correlation plot between principle coordinate axis 1 and death rate of isolates in FPW

Partial Least Square (PLS) regression analysis of the number of days required to recover culturability in the APW microcosms using a variable importance threshold (VIP) of 0.8 reduced the number of genes under consideration from 1288 to 837. PLS regression

analysis of a strain's death rate in the FPW microcosms using a VIP of 0.8 resulted in 596 of the 1288 genes being retained.

Fit model analysis revealed that three genes, *ariR*, *gatA*, and group_40 (hypothetical protein) explained a significant amount of the variation in the number of days required for a strain to regain culturability in the APW microcosms (Table 4.7).

Table 4.7 The results of fit model analysis to determine which genes of an isolate's variable gene pool best explained the number of days taken to regain culturability after entering the VBNC state in APW

Source	LogWorth	P Value
ariR	2.123	0.007
gatA_2	1.759	0.017
group_40	1.234	0.058
yeeJ_2	0.966	0.108
group_5005	0.692	0.203
nadB_1	0.466	0.342
higB-1	0.135	0.733
mrr	0.085	0.822
group_2867	0.067	0.856
imm_2	0.03	0.932

The gene *ariR*, functions as a regulator for acid resistance (Lee et al., 2007), while gatA is a galactitol transfer phosphotransferase protein (Noblemann and Lengeler, 1996; Volpon et al., 2006), and the gene designated group_40 has been annotated as a host specificity protein. The nucleotide sequences of the genes are presented in Appendix D.1. Individually, the presence of either *ariR* or group_40 in a strain decreases the number of days taken by a strain to return to culturability, while the presence of *gatA* increases the number of days required.

The presence/absence of these three genes among the 50 strains resulted in six profiles. Strains with *ariR* and group_40, but lacking *gatA* took the fewest days to recovery

culturability (6 days), while those strains with both *ariR* and *gatA*, but lacking group_40 took significantly longer to recover (101days) (Table 4.8).

Table 4.8 The presence/absence profile of the three genes (Table 4.7) explaining the most variation in the number of days taken to regain culturability after entering the VBNC state in APW

Genes			**Connecting letter p<0.05**	**Number of isolates**	**Mean Days to recovery**	**Effect on the days to recovery**
ariR	***gatA***	***group_40***				
-	-	+	A B	1	111	Increase
+	+	-	A	11	101	Increase
-	-	-	A B	4	73	Increase
+	+	+	B	7	57	Decrease
+	-	-	B	7	35	Decrease
+	-	+	C	7	6	Decrease

Fit model analysis identified the three most important genes in explaining the variation in a strain's death rate in FPW, and these were found to be *yfdM*, *pnuC*, *group_1212* (Table 4.9).

Table 4.9 The results of fit model analysis to determine which genes of an isolate's variable gene pool best explained the variation in the death rate of isolates in FPW.

Source	LogWorth	P Value
yfdM	1.909	0.012
pnuC_2	1.458	0.034
group_1212	1.341	0.045
group_954	1.144	0.071
group_1105	0.834	0.146
group_2125	0.573	0.267
yfdM_1	0.373	0.423
nadR_2	0.356	0.440
group_1274	0.139	0.725
group_236	0.058	0.874
group_5733	0.045	0.900

The presence of the genes *pnuC,* a nicotinamide riboside transfer salvage protein (Sauer et al., 2004), and group_1212, a hypothetical protein increases the death rate of isolates in FPW, while the presence of the gene *yfdM*, a putative methyl transferase protein, decreases a strains death rate in FPW. The nucleotide sequences of the genes are presented in Appendix D.2

The presence/absence of these three genes among the 50 strains resulted in eight profiles. The presence of two of these genes in a strain increased the death rate of the strains, while the highest death rates were observed for strains having all three genes (Table 4.10).

Table 4.10 The presence/absence profile of the three genes (Table 4.9) found to explain the most variation in an isolate's death rate in filtered pond water (FPW).

Genes			Connecting letter, $p<0.05$	Number of isolates	Mean Death rate (day^{-1})	Effect on death
group_1212	***pnuC***	***yfdM***				
+	+	+	A B C D	2	0.043	Increase
-	+	+	B	4	0.017	Increase
+	+	-	B C	4	0.015	Increase
+	-	+	B C D	1	0.014	Increase
-	+	-	C D	7	0.005	Decrease
-	-	+	B C D	4	0.005	Decrease
-	-	-	D	24	0.004	Decrease
+	-	-	B C D	2	0.004	Decrease

Discussion

Investigating the survival of *E. coli* in external environments can provide valuable insights into its ecological niche, survival strategies, and environmental impacts especially in relation to food and water safety. While the persistence of *E. coli* in human hosts (Clermont et al., 2008; Katouli, 2010; Gordon et al., 2015) and *E. coli* as a human pathogen (Kaper et al., 2004; Chandran and Mazumder, 2015; Jang et al., 2017) have been extensively studied, a few studies have explored on the survival of *E. coli* in external environments. Phylogroup B2 strains are thought to be the most host adapted of the various *E. coli* phylogroups, but have also been thought to be the least capable of surviving in water (Walk et al., 2007; Berthe et al., 2013; Quero et al., 2015). This study investigated the survival of 50 strains of the B2 phylogroup isolated from water and vertebrates, including humans.

The microcosm study was conducted using three different water treatments, sterilised deionised water, autoclaved and filter sterilised pond water. Although various factors affect the survival of *E. coli* in external environments such as pH, salinity, sunlight intensity, and predation, temperature is the major factor determining variation in the rate of cell division and survival (Rigsbee et al., 1997; Presser et al., 1998; Faust et al., 1975; Flint, 1987; Bordalo et al., 2002; Sinton et al., 2002; McCambridge and McMeekin, 1980; Kudva et al., 1998; Blaustein et al., 2013; Wanjugi et al., 2016). Since most of the water bodies in Australia maintain a median temperature between 18°C and 22°, the microcosm experiment was conducted at 20°C. This experiment was also carried out under a constant pH (7.0) environment and in the absence of other organisms such as protozoa. *E. coli* cell densities in natural aquatic environments are typically $<10^3$/ml even during outbreak events (Fewtrell et al., 1994; Olsen et al., 2002; Licence et al., 2001). Hence this microcosm experiment used a starting density of 10^3 CFUs/ml. As the experiment proceeded, isolates in this experiment exhibited a diverse range of survival patterns with respect to the water microcosm variations tested.

The deionised water microcosms represented an environment free of any nutrients and one that would subject the *E. coli* cells to a high osmotic stress. Cells in this environment died rapidly and were apparently unable to enter a VBCN state.

In APW, some phylogroup B2 strains are clearly capable of entering the VBNC state and spontaneously returning to culturablity within 2 to > 90 days. Why was the VBNC state observed in the APW but not FPW microcosms? The results of this study indicate that the heat treatment caused the production of ammonia from organic compounds present in the pond water. Similar outcomes were observed by Wang and Doyle, 1997 and Liu et al., 2009, where autoclaved municipal and river water induced a VBNC state in *E. coli* at 7 and 14 weeks, respectively. Also, Yeung et al., 2006 reported that autoclave sterilisation of water mixed with powdered infant formula increased the level of ammonia due to a Maillard reaction that in turn degraded many proteins and free amino acid constituents in the milk. Other studies have also suggested that elevated levels of ammonia have negative influence on the growth of *E. coli* (Niebuhr et al., 2003; Park et al., 2003).

The time taken to return to culturability has a genetic component, but what causes the cells to spontaneously leave the VBNC state is unknown. Two recent studies by Kim and collegues (2018 a & b) examined VBNC cells and their resuscitation rate. In the first study, they induced 10^8 *E. coli* cells/ml into a VBNC state through long-term nutrient depletion and attempted resuscitation by the addition of growth promoting factors. They found that only cells that were rod shaped with a cytosol content could be successfully resuscitated, and that these cells represented a small fraction of the total cell population. Cells with no cytosol content represented dead cells. In the second study they determined that resuscitated cells either initiated cell division immediately or showed a delayed start to cell division. They found that cells initiating replication immediately had four-fold higher ribosomal content prior to entering VBNC state than the cells that had delayed cell division.

The results from the FPW microcosms clearly demonstrate that phylogroup B2 *E. coli* isolates are physiologically capable of surviving for extended periods without entering the VBNC state. The average lifespan of cells in water also appears to have a genetic component. The finding of long-time survival in the filtered water microcosms is surprising, as any nutrients available at the start of the study would eventually have been consumed, yet there was no indication that a significant fraction of the cells present entered a VBCN state.

Artz and Killham, 2002 studied the survival at 15°C of *E. coli* 0157:H7 at a concentration of 10^9 CFU/ml in waters from four private drinking water wells in Scotland. Water from wells 1 & 2 was considered to be of higher quality than water from wells 3 & 4, which had high concentrations of heavy metals. The water from all four wells was treated in one of four ways (i) untreated (ii) filtered using 3µm filter (iii) filtered using 0.2µm filter (iv) autoclaved. In their experiment the 0157:H7 isolate survived substantially longer in 0.2µm filtered water and autoclaved water from wells 1 & 2 compared to its survival in untreated or 3µm filtered water. The authors attributed these outcomes to the absence of predators in water that had been autoclaved or filtered through a 0.2 µm filter. However, in all the microcosms using water from wells 1 & 2 there was an initial rapid, about 100-fold, decline in the number of CFU/ml in the first 3-4 days of the experiment. After this initial decline, the number of CFU/ml remained relatively constant over the 65 days of sampling when predators were absent, but eventually declined to zero in the microcosms where predators were present. Conversely, the survival of the O157:H7 isolate was very much poorer in the water contaminated with heavy metals regardless of the water treatment. Measures of total viable cells indicated that the observed declines, regardless of treatment or water source, were due to cell death and not due to cells entering the VBNC state. The fact that there was no evidence of cells entering the VBNC state even in the autoclave treated water samples, suggests that the source of the water that is subsequently autoclaved are important. Compared to the pond water used in the present experiments, it is likely that the well water would have lower levels of dissolved organic matter.

Similar outcomes were observed in two studies by Wanjugi and Harwood, 2014 and Wanjugi et al., 2016. In the first study, the survival of motile vs. non-motile *E. coli* O157 was investigated in microcosms with and without predators. The microcosms had a starting cell density of 10^8 CFU/100ml and the predation treatment was affected by introducing the protozoan *Tetrahymena pyriformis*. They found that CFU/microcosm decreased rapidly in the presence of predation, but in the absence of predation, the CFU/ml increased or declined very slowly for both motile and non-motile *E. coli* O157. In the second study the authors assessed the importance of natural nutrients, natural predators and bacterial competition on the survival of *E. coli* by setting up four different microcosms each having nutrients such as glucose, pyruvate, acetate, trace elements and minerals added at three nutrient levels of 0x (nil), 1x, 5x. Each microcosm was inoculated

with 10^8 CFU/ 100ml *E. coli* in fresh river water. Treatment one retained the natural predators and competitors, while for treatment two the indigenous protozoa and bacteria were removed by filter sterilisation. For treatment three kanamycin was added to inhibit the growth of indigenous bacteria, leaving the protozoan predators, while in treatment four the addition of cycloheximide inhibited the indigenous protozoa, leaving the bacteria. They found that only in treatment two (no predation or competition) did the CFU/ microcosm increased with increase in nutrient level. In the balance of the treatments CFU/ microcosm declined. Thus, in the present study, the absence of predators coupled with the nutrient rich pond water that was used, likely explains the excellent survival of isolates in the FPW microcosms.

In this study, *in silico* analyses of the variable gene content of the B2 isolates clearly suggested that gene content plays a role in the time taken to recover in APW and their survival in FPW. In APW, although the genes *ariR*, a hypothetical protein group_40 and *gatA* are the top three genes predicted to cause variation in cells recovery from VBNC state, it is not obvious why an acid regulating gene *ariR* would influence recovery time in a stable pH environment, or why the galactitol transfer protein *gatA* would increase the number of days cells take to recover from VBNC. Similarly, in FPW, why a nicotinamide riboside transfer salvage protein gene *pnuC* increases, while the putative methyl transferase gene *yfdM* decreases the survival of isolates is also unknown and requires further research.

Overall, the APW and FPW microcosms demonstrate what phylogroup B2 strains are capable of in terms of their survival in aquatic habitats. However, these experiments represent ideal conditions for *E. coli*'s survival.

What are the implications of the outcomes for the use of *E. coli* as a FIB in the real world? It is evident from this study that *E. coli* can enter the VBNC state or survive long term in aquatic environments. In raw and treated drinking water contaminated with *E. coli*, if ammonia is provided by chloramination treatment, it may induce *E. coli* into VBNC state. Cells entering the VBNC state will not be detected by the standard methods used to determine *E. coli* counts in water. If a significant fraction of cells enters the VBNC state immediately following a faecal contamination event, the event will be missed along with the pathogenic *E. coli* and other faecal pathogens that might have also been introduced. The VBNC state of *E. coli* may cause a risk to human health if the virulence associated

genes are still producing toxins and actively transmit virulence after recovery (Schottroff et al., 2018; Pienaar et al., 2016; Makino et al., 2000; Liu et al., 2010; Dinu and Bach 2011; Oliver 2000;). Although the VBNC state could certainly have an impact, there is no concrete evidence showing that a significant fraction of the *E. coli* being deposited in natural aquatic environments become VBNC. Similarly, although nutrient limitation, low temperature and pH changes have been shown to induce VBNC in *E. coli*, it is not clear if nutrient limitations and or changing 'natural' aquatic environmental conditions is sufficient to induce the VBNC state. Further, for *E. coli* to recover from a VBNC state or to grow to a large number, it requires no predation and elevated nutrient levels. However, raw water is highly unlikely to be absent of predators and in treated drinking water, although there will be no or low number of predators, *E. coli* numbers will be controlled by disinfectants. Potentially, the average lifespan of *E. coli* in real habitats will be much shorter than observed in the present study, but how much shorter will depend on the local predation pressure and other environmental variables that adversely impact survival.

The outcome of this study is in contrast to the experimental finding of Berthe et al., (2013) comparing the survival of strains belonging to phylogroups A, B1, B2 in estuarine waters in France. Their experiments were conducted at 10°C with 10^7 CFU/ml initial cell density for 14 days. They found that phylogroup B1 and to a lesser extend phylogroup A strains survived for more than 14 days and phylogroup B2 and D survived for less than 4 days. It may be that the survival of strains belonging to the different phylogroups differ in their temperature responses. A study of the thermal niche of *E. coli* isolates by Okada and Gordon (2001) provides some support for this hypothesis. They observed that phylogroup B1 isolates could maintain their stationary phase cell density in glucose-limited serial transfer experiments for temperatures between 17-18°C, by contrast the temperatures at which B2 isolates could do this ranged from 17- 22°C.

Given, that B2 strains are physiologically capable of surviving for extended periods in aquatic habitats, why are B2 strains typically one of the least abundant phylogroups to be observed in water samples? (Picard et al., 1999; Power et al., 2005; Walk et al., 2007; Touchon et al., 2009). If B2 isolates do not respond differently to temperature compared to strains of the other phylogroups, then perhaps differential predation pressure might explain the low abundance of B2 isolates in water compared to isolates of the other

phylogroups. Studies with *Salmonella* have shown that a strains susceptibility to predation varies with its serotype (Atzinger et al., 2016) and in *E. coli* particular serotypes are over-represented in B2 strains compared to strains of the other phylogroups (Johnson and Stell, 2000; O'Brien et al., 2016). Further research is required to validate this prediction.

Finally, the relative rarity of B2 isolates in natural aquatic environments may simply be a reflection of their rarity in the faecal inputs to such environments. However, there is little evidence to support this hypothesis, as B2 strains represent about 35% of the *E. coli* recovered from Australian mammals and 22% of isolates from Australian birds (Gordon & Cowling, 2005).

Overall, this study suggests the source (host versus environment) of *E. coli* isolates and differences in water sterilisation treatments both play a significant role in *E. coli's* adaption and survival in the external environment. Further, the findings that *E. coli* isolates not only persisted for a long time but also may become VBNC cast doubt on the effectiveness of *E. coli* as a reliable indicator of recent faecal contamination.

Reference

Abberton, C. L., Bereschenko, L., van der Wielen, P. W. J. J. & Smith, C. J., 2016. Survival, Biofilm Formation, and Growth Potential of Environmental and Enteric *Escherichia coli* Strains in Drinking Water Microcosms. *Applied and Environmental Microbiology,* 82(17), pp. 5320-5331.

Anderson, M. et al., 2004. Viable but Nonculturable Bacteria Are Present in Mouse and Human Urine Specimens. *Journal of Clinical Microbiology,* 42(2), pp. 753-758.

Arana, I. et al., 2007. Inability of *Escherichia coli* to resuscitate from the viable but nonculturable state. *FEMS Microbiology Ecology,* 62(1), pp. 1-11.

Artz, R. R. & Killham, K., 2002. Survival of *Escherichia coli* O157:H7 in private drinking water wells: influences of protozoan grazing and elevated copper concentrations. *FEMS Microbiology Letters,* 216(1), pp. 117-122.

Asakura, A. et al., 2007. Increased survival of muscle stem cells lacking the MyoD gene after transplantation into regenerating skeletal muscle. *Proceedings of the National Academy of Sciences,* 104(42), pp. 16552-16557.

Atzinger, A., Butelat, K. & Lawrence, J. G., 2016. The O-antigen mediates differential survival of Salmonella against communities of natural predators. *Microbiology,* Volume 162, pp. 610-621.

Barcina, I., Gonzalez, J. M., Iriberri, J. & Egea, L., 1990. Survival strategy of *Escherichia coli* and Enterococcus faecalis in illuminated fresh and marine systems. *The Journal of Applied Bacteriology,* 68(2), pp. 189-198.

Berthe, T. et al., 2013. Evidence for Coexistence of Distinct *Escherichia coli* Populations in Various Aquatic Environments and Their Survival in Estuary Water. *Applied and Environmental Microbiology,* 79(15), p. 4684–4693.

Blackburn, C. W. & McCarthy, J. D., 2000. Modifications to methods for the enumeration and detection of injured *Escherichia coli* O157:H7 in foods. *International Journal of Food Microbiology,* 55((1-3)), pp. 285-290.

Blaustein, R. A. et al., 2013. *Escherichia coli* survival in waters: Temperature dependence. *Water Research,* 47(2), pp. 569-578.

Blyton, M. D. J. & Gordon, D. M., 2017. Genetic Attributes of *E. coli* Isolates from Chlorinated Drinking Water. *PLoS ONE,* 12(1), p. e0169445.

Blyton, M. D. et al., 2015. Genetic Structure and Antimicrobial Resistance of *Escherichia coli* and Cryptic Clades in Birds with Diverse Human Associations. *Applied and Environmental Microbiology,* 81(15), p. 5123–5133.

Boehm, A. et al., 2009. Second messenger signalling governs *Escherichia coli* biofilm induction upon ribosomal stress. *Molecular Microbiology,* 72(6), pp. 1500-1516.

Bordalo, A. A., Onrassami, R. & Dechsakulwatana, C., 2002. Survival of faecal indicator bacteria in tropical estuarine waters (Bangpakong River, Thailand). *Journal of Applied Microbiology,* 93(5), pp. 864-871.

Byrd, J. J. & Colwell, R. R., 1993. Long-term survival and plasmid maintenance of *Escherichia coli* in marine microcosms. *FEMS Microbiology Ecology,* 12(1), pp. 9-14.

Cappelier, J. M. et al., 2007. Avirulent Viable But Non Culturable cells of *Listeria monocytogenes* need the presence of an embryo to be recovered in egg yolk and regain virulence after recovery. *EDP Sciences,* Volume 38, pp. 573-583.

Castro Stoppe, N. d. et al., 2017. Worldwide Phylogenetic Group Patterns of *Escherichia coli* from Commensal Human and Wastewater Treatment Plant Isolates. *Frontiers in Microbiology,* 8(2512).

Chandran, A. & Mazumder, A., 2015. Pathogenic Potential, Genetic Diversity, and Population Structure of *Escherichia coli* Strains Isolated from a Forest-Dominated Watershed (Comox Lake) in British Columbia, Canada. *Applied and Environmental Microbiology,* 81(5), p. 1788 –1798.

Clermont, O., Christenson, J. K., Denamur, E. & Gordon, D. M., 2013. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports,* 5(1), pp. 58-65.

Clermont, O. et al., 2008. Evidence for a human-specific *Escherichia coli* clone. *Environmental Microbiology,* 10(4), pp. 1000-1006.

Coli, D., Jospin, G. & Darling, A. E., 2015. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics (Oxford, England)*, 31(4), pp. 587-589.

Colwell, R. R., 2009. Viable but Not Cultivable Bacteria. In: S. S. Epstein, ed. *Uncultivated Microorganisms*. Berlin, Heidelberg: Springer, pp. 121-129.

Cuny, C. et al., 2005. Investigation of the First Events Leading to Loss of Culturability during *Escherichia coli* Starvation: Future Nonculturable Bacteria Form a Subpopulation. *Journal of Bacteriology,* 187(7), pp. 2244-2248.

Ding, T. et al., 2017. Significance of Viable but Nonculturable *Escherichia coli*: Induction, Detection, and Control. *Journal of Microbiology and Biotechnology*, 27(3), pp. 417-423.

Dinu, L. D. & Bach, S., 2011. Induction of viable but nonculturable *Escherichia coli* O157:H7 in the phyllosphere of lettuce: a food safety risk factor. *Applied and Environmental Microbiology,* 77(23), pp. 8295-8302.

Dixit , O. V. A., O'Brien, C. L., Pavli, P. & Gordon, D., 2018. Within-host evolution versus immigration as adeterminant of *Escherichia coli* diversity in the human gastrointestinal tract. *Environmental Microbiology ,* 20(3), pp. 993-1001.

Draper, J. et al., 2004. *Metabolite Peak Identification and Data Structure in a Multi-Site, Large Scale Metabolomics Experiment.* s.l., PittCon.

Escobar-Páramo, P. et al., 2006. Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates. *Environmental Microbiology,* 8(11), pp. 1975-1984.

Faust, M. A., Aotaky, A. E. & Hargadon, M. T., 1975. Effect of Physical Parameters on the In Situ Survival of *Escherichia coli* MC-6 in an Estuarine Environment. *Applied Microbiology,* 30(5), pp. 800-806.

Fewtrell, L. et al., 1994. The Health Effects of Low-Contact Water Activities in Fresh and Estuarine Waters. *Water and Environment Journal,* 8(1), pp. 97-101.

Flint, K. P., 1987. The long-term survival of *Escherichia coli* in river water. *The Journal of Applied Bacteriology,* 63(3), pp. 261-270.

Gordon, D. M., 2001. Geographical structure and host specificity in bacteria and the implications for tracing the source of coliform contamination. *Microbiology,* Volume 147, pp. 1079-1085.

Gordon, D. M. & Cowling, A., 2003. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology,* Volume 149, p. 3575–3586.

Gordon, D. M., O'Brien, C. L. & Pavli, P., 2015. *Escherichia coli* diversity in the lower intestinal tract of humans. *Environmental Microbiology Reports,* 7(4), pp. 642-648.

Gordon, D. M., Stern, S. E. & Collignon, P. J., 2005. Influence of the age and sex of human hosts on the distribution of *Escherichia coli* ECOR groups and virulence traits. *Microbiology,* 151(1), pp. 15-23.

Hartl, D. L. & Dykhuizen, D. E., 1984. The population genetics of *Escherichia coli*. *Annual Review of Genetics* , Volume 18, pp. 31-68.

Harwood, V. J. et al., 2014. Microbial source tracking markers for detection of fecal contamination in environmental waters: relationships between pathogens and human health outcomes. *FEMS Microbial Reviews,* Volume 38, pp. 1-40.

Ihssen, J. & Egli, T., 2005. Global physiological analysis of carbon- and energy-limited growing *Escherichia coli* confirms a high degree of catabolic flexibility and preparedness for mixed substrate utilization. *Environmental Microbiology,* 7(10), pp. 1568-1581.

Jang, J. et al., 2017. Environmental *Escherichia coli*: ecology and public health implications-a review. *Journal of Applied Microbiology,* 123(3), pp. 570-581.

Johnson, J. R. & Stell, A. L., 2000. Extended Virulence Genotypes of *Escherichia coli* Strains from Patients with Urosepsis in Relation to Phylogeny and Host Compromise. *The Journal of Infectious Diseases,* 181(1), pp. 261-272.

Jones, B. & Sall, J., 2011. JMP statistical discovery software. *WIREs Computational Statistics,* 3(3), pp. 188-194.

Juhna, T. et al., 2007a. Detection of *Escherichia coli* in Biofilms from Pipe Samples and Coupons in Drinking Water Distribution Networks. *Applied and Environmental Microbiology,* 73(22), pp. 7456-7464.

Juhna, T., Birzniece, D. & Rubulis, J., 2007b. Effect of Phosphorus on Survival of *Escherichia coli* in Drinking Water Biofilms. *Applied and Environmental Microbiology,* 73(11), p. 3755–3758.

Kaper, J. B., Nataro, J. P. & Mobley, H. L., 2004. Pathogenic *Escherichia coli*. *Nature Reviews Microbiology,* 2(2), pp. 123-140.

Katouli, M., 2010. Population structure of gut *Escherichia coli* and its role in development of extra-intestinal infections. *Iranian Journal of Microbiology,* 2(2), pp. 59-72.

Keer, J. T. & Birch, L., 2003. Molecular methods for the assessment of bacterial viability. *Journal of Microbiological Methods,* Volume 53, pp. 175-183.

Kim, J. S., Chowdhury, N., Yamasaki, R. & Wood, T. K., 2018a. Viable but non-culturable and persistence describe the same bacterial stress state. *Environmental Microbiology,* 20(6), pp. 2038-2048.

Kim, J. S. et al., 2018b. Single cell observations show persister cells wake based on ribosome content. *Environmental Microbiology,* 20(6), pp. 2085-2098.

Klein, T. M. & Alexander, M., 1986. Bacterial Inhibitors in Lake Water. *Applied and Environmental Microbiology,* 52(1), pp. 114-118.

Kudva, I. T., Blanch, K. & Hovde, C. J., 1998. Analysis of *Escherichia coli* O157:H7 survival in ovine or bovine manure and manure slurry. *Applied and Environmental Microbiology,* 64(9), pp. 3166-3174.

Le Gall, T. et al., 2007. Extraintestinal Virulence Is a Coincidental By-Product of Commensalism in B2 Phylogenetic Group *Escherichia coli* Strains. *Molecular Biology and Evolution,* 24(11), pp. 2373-2384.

Leclerc, H., Mossel, D. A., Edberg, S. C. & Struijk, C. B., 2001. Advances in the bacteriology of the coliform group: their suitability as markers of microbial water safety. *Annual Review of Microbiology,* Volume 55, pp. 201-234.

Lee, J. et al., 2007. Structure and function of the *Escherichia coli* protein YmgB: a protein critical for biofilm formation and acid-resistance.. *Journal of Molecular Biology* , 373(1), pp. 11-26.

Licence, K., Oates, K. R., Synge, B. A. & Reid, T. M., 2001. An outbreak of *E. coli* O157 infection with evidence of spread from animals to man through contamination of a private water supply. *Epidemiology and Infection,* 126(1), pp. 135-138.

Li, L. et al., 2014. The importance of the viable but non-culturable state in human bacterial pathogens. *Frontiers in Microbiology,* 5(258).

Liu, Y., Gilchrist, A., Zhang, J. & Li, X. F., 2008. Detection of Viable but Nonculturable *Escherichia coli* O157:H7 Bacteria in Drinking Water and River Water. *Applied and Environmental Microbiology,* 74(5), pp. 1502-1507.

Liu, Y. et al., 2009. Induction of *Escherichia coli* O157:H7 into the viable but non-culturable state by chloraminated water and river water, and subsequent resuscitation. *Environmental Microbiology Reports,* 1(2), pp. 155-161.

Liu, Y., Wang, C., Tyrrell, G. & Li , X. F., 2010. Production of Shiga-like toxins in viable but nonculturable *Escherichia coli* O157:H7. *Water Research,* 44(3), pp. 711-718.

Makino, S. I. et al., 2000. Does enterohemorrhagic *Escherichia coli* O157:H7 enter the viable but nonculturable state in salted salmon roe? *Applied and Environmental Microbiology,* 66(12), pp. 5536-5539.

McCambridge, J. & McMeekin, T. A., 1980. Relative effects of bacterial and protozoan predators on survival of *Escherichia coli* in estuarine water samples. *Applied and Environmental Microbiology,* 40(5), pp. 907-911.

McKay, A. M., 1992. Viable but non-culturable forms of potentially pathogenic bacteria in water. *Letters in Applied Microbiology,* 14(4), pp. 129-135.

Muela, A. et al., 2008. Changes in *Escherichia coli* outer membrane subproteome under environmental conditions inducing the viable but nonculturable state. *FEMS Microbiology Ecology,* 64(1), pp. 28-36.

Na, S. H., Miyanaga, K., Unno, H. & Tanji, Y., 2006. The survival response of *Escherichia coli* K12 in a natural environment. *Applied Microbiology and Biotechnology,* 72(2), pp. 386-392.

Niebuhr, S. E. & Dickson, J. S., 2003. Impact of pH enhancement on the populations of Salmonella, Listeria and *Escherichia coli* O157:H7 in boneless lean beef trimmings. *Journal of Food Protection,* 66(5), pp. 874-877.

Nobelmann, B. & Lengeler, J. W., 1996. Molecular analysis of the *gat* genes from *Escherichia coli* and of their roles in galactitol transport and metabolism. *Journal of Bacteriology,* 178(23), pp. 6790-6795.

Nowrouzian, F. L., Adlerberth, I. & Wold, A. E., 2006. Enhanced persistence in the colonic microbiota of *Escherichia coli* strains belonging to phylogenetic group B2: role of virulence factors and adherence to colonic cells. *Microbes and Infection,* 8(3), pp. 834-840.

O'Brien, C. L. et al., 2016. Comparative genomics of Crohn's disease-associated adherent-invasive *Escherichia coli*. *Gut Microbiota,* Volume doi:10.1136/gutjnl- 2015-311059, pp. 1-8.

Odonkor, S. T. & Ampofo, J. K., 2013. *Escherichia coli* as an indicator of bacteriological quality of water: an overview. *Microbiology research,* 4(1), p. e2.

Oliver, J. D., 2000. The viable but nonculturable state and cellular resuscitation. In: C. R. Bell, M. Brylinsky & P. Johnson-Green, eds. *Microbial Biosystems: New Frontiers.* Halifax, Canada: Atlantic Canada Society for Microbial Ecology, pp. 723-730.

Oliver, J. D., 2005. The Viable but Nonculturable State in Bacteria. *The Journal of Microbiology,* 43(5), pp. 93-100.

Olsen, S. J. et al., 2002. A Waterborne Outbreak of *Escherichia coli* O157:H7 Infections and Hemolytic Uremic Syndrome: Implications for Rural Water Systems. *Emerging Infectious Diseases,* 8(4), pp. 370-375.

Page, A. J. et al., 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics,* 31(22), pp. 3691-3693.

Park, G. W. & Diez-Gonzalez, F., 2003. Utilization of carbonate and ammonia-based treatments to eliminate *Escherichia coli* O157:H7 and Salmonella Typhimurium DT104 from cattle manure. *Journal of Applied Microbiology,* 94(4), pp. 675-685.

Picard, B. et al., 1999. The Link between Phylogeny and Virulence in *Escherichia coli* Extraintestinal Infection. *Infection and Immunity,* 67(2), pp. 546-553.

Pienaar, J. A., Singh, A. & Barnard, T. G., 2016. The viable but non-culturable state in pathogenic *Escherichia coli*: A general review. *African Journal of Laboratory Medicine,* 5(1), p. 368.

Pinto, D., Almeida, V., Almeida Santos, M. & Chambel, L., 2011. Resuscitation of *Escherichia coli* VBNC cells depends on a variety of environmental or chemical stimuli. *Journal of Applied Microbiology,* 110(6), pp. 1601-1611.

Power, M. L. et al., 2005. Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. *Environmental Microbiology,* 7(5), pp. 631-640.

Presser, K. A., Ross, T. & Ratkowsky, D. A., 1998. Modelling the growth limits (growth/no growth interface) of *Escherichia coli* as a function of temperature, pH, lactic acid concentration, and water activity. *Applied and Environmental Microbiology*, 64(5), pp. 1773-1779.

Quero, G. M., Fasolato, L., Vignaroli, C. & Luna, G. M., 2015. Understanding the association of *Escherichia coli* with diverse macroalgae in the lagoon of Venice. *Scientific Reports*, Volume 5, p. 5:10969.

Reissbrodt, R. et al., 2002. Resuscitation of Salmonella enterica Serovar Typhimurium and Enterohemorrhagic *Escherichia coli* from the Viable but Nonculturable State by Heat-Stable Enterobacterial Autoinducer. *Applied and Environmental Microbiology*, 68(10), p. 4788–4794.

Rigsbee, W., Simpson, L. M. & Oliver, J. D., 1997. Detection of the viable but nonculturable state in *Escherichia coli* O157:H7. *Journal of Food Safety*, 16(4), pp. 255-262.

Sachidanandham, R. & Gin, Y. H. K., 2009. A dormancy state in nonspore-forming bacteria. *Applied Microbiology and Biotechnology*, 81(5), pp. 927-941.

Sauer, E. et al., 2004. *PnuC* and the utilization of the nicotinamide riboside analog 3-aminopyridine in Haemophilus influenzae. *Antimicrobial Agents and Chemotherapy*, 48(132), pp. 4532-4541.

Savageau, 1983. *Escherichia coli* habitats, cell types, and molecular mechanisms of gene control. *The American Naturalist*, 122(6), pp. 732-744.

Schottroff, F. et al., 2018. Sublethal Injury and Viable but Non-culturable (VBNC) State in Microorganisms During Preservation of Food and Biological Materials by Non-thermal Processes. Frontiers in Microbiology, 9(2773), p. doi: 10.3389/fmicb.2018.02773.

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14), pp. 2068-2069.

Sinton, L. W., Hall, C. H., Lynch, P. A. & Davies-Colley, R. J., 2002. Sunlight inactivation of fecal indicator bacteria and bacteriophages from waste stabilization pond effluent in fresh and saline waters. *Applied and Environmental Micrbiology,* 68(3), pp. 1122-1131.

Smati, M. et al., 2015. Quantitative analysis of commensal *Escherichia coli* populations reveals host-specific enterotypes at the intra-species level. *Microbiology,* 4(4), pp. 604-615.

R Core Team. *R: A language and Environment for Statistical Computing.* [Online] Available at: https://www.R-project.org [Accessed 2017].

Touchon, M. et al., 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLOS Genetics,* 5(1), p. 5:e1000344.

van Elsas, J. D., Semenov, A. V., Costa, R. & Trevors, J. T., 2011. Survival of *Escherichia coli* in the environment: fundamental and public health aspects. *The ISME Journal,* Volume 5, pp. 173-183.

Versalovic, J., Koeuth, T. & Lupski, J. R., 1991. Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. *Nucleic Acid Research,* 19(24), p. 6823–6831.

Vital, M., Hammes, F. & Egli, T., 2008. *Escherichia coli* O157 can grow in natural freshwater at low carbon concentrations. *Environmental Microbiology,* 10(9), pp. 2387-2396.

Volpon, L., Young, C. R., Matte, A. & Gehring, K., 2006. NMR structure of the enzyme GatB of the galactitol-specific phosphoenolpyruvate-dependent phosphotransferase system and its interaction with GatA. *Protein Structure Report,* Volume 15, pp. 2435-2441.

Walk, S. T. et al., 2007. Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. *Environmental Microbiology,* 9(9), pp. 2274-2288.

Wang, G. & Doyle, M. P., 1998. Survival of Enterohemorrhagic *Escherichia coli* O157:H7 in Water. *Journal of food protection*, 61(6), pp. 662-667.

Wanjugi, P., Fox, G. A. & Harwood, V. J., 2016. The Interplay Between Predation, Competition, and Nutrient Levels Influences the Survival of *Escherichia coli* in Aquatic Environments. *Microbiology of Aquatic Systems*, Volume 72, pp. 526-537.

Wanjugi, P. & Harwood, V. J., 2014. Protozoan Predation Is Differentially Affected by Motility of Enteric Pathogens in Water vs. Sediments. *Environmental Microbiology*, Volume 68, pp. 751-760.

Wanjugi, P. et al., 2016. Differential decomposition of bacterial and viral fecal indicators in common human pollution types. Water Research, Volume 105, pp. 591-601.

Ward, D. M., Weller, R. & Bateson, M. M., 1990. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature*, 345(6270), pp. 63-65.

WHO, 1993. *Guidelines for Drinking Water Quality*. Second ed. World Health Organization, Geneva: Volume 2 Health criteria and other supporting information.

Xu, H. S. et al., 1982. Survival and Viability of Nonculturable *Escherichia coli* and *Vibrio cholerae* in the Estuarine and Marine Environment. *Microbial Ecology*, Volume 8, pp. 313-323.

Yeung , C. Y. et al., 2006. Negative effect of heat sterilization on the free amino acid concentrations in infant formula. *European Journal of Clinical Infection*, 60(1), pp. 136-141.

Zhang, S. et al., 2015. UV disinfection induces a VBNC state in *Escherichia coli* and Pseudomonas aeruginosa. *Environmental Science and Technology*, 49(3), pp. 1721-1728.

Zhao, F., Bi, X., Hao, Y. & Liao, X., 2013. Induction of Viable but Nonculturable *Escherichia coli* O157:H7 by High Pressure CO_2 and Its Characteristics. *PLoS ONE*, 8(4), p. e62388.

Zhao, F. et al., 2016. New Insights into the Formation of Viable but Nonculturable *Escherichia coli* O157:H7 Induced by High-Pressure CO_2. *mBio*, 7(4), pp. e00961-16.

CHAPTER 5

Conclusions

This thesis investigated the limitations of using *E. coli* as an effective indicator for faecal contamination, particularly recent faecal contamination of water bodies. Universally, water safety guidelines use *E. coli* as a major indicator for recent faecal contamination in aquatic environments (EPA, 1986; WHO, 2008). This is mainly because of the assumption that *E. coli* is present at a high concentration in mammalian gut and faeces, and when it is discharged into the external environment such as soil and water, the *E. coli* rapidly dies, surviving for less than 5 days (Hartl and Dykhuizen, 1984; Slanetz and Bartley, 1957; Faust et al., 1975; Ingle et al., 2011). Consequently, it is considered as a transient member of the natural environment reflecting the same clonal composition as that of the source population responsible for faecal input, and reflecting possible presence of other faecal pathogens thought to be deposited with it. Recently, the reliance on *E. coli* as a FIB is being more closely scrutinised as an increasing number of studies detect *E. coli*'s presence in the external environment without any faecal contamination (Power et al., 2005; Walk et al., 2007; Brennan et al., 2010). Hence, water authorities want a quick and inexpensive way to assess whether water has been contaminated with faeces.

To examine whether a subset of *E. coli* that are likely to have suitable characteristics should be used as a FIB, the phylogroup B2 was studied in this thesis since: a) it is the most host associated phylogroup and specialized to vertebrate isolates, especially those with large hind gut fermentation b) it is more commonly isolated from humans than other phylogroups of *E. coli*, mainly in industrialised countries like Australia (Gordon and Cowling, 2003; Gordon et al., 2005; Blyton et al., 2015; Gordon et al., 2015), c) *E. coli* from human faeces is considered to be the highest risk category of faecal pollution in water as it possess greater risk to human health due to the diseases associated with it (Regli et al., 1991; Harwood et al., 2014) especially phylogroup B2, which has more virulence factors to adhere and cause diseases than any other phylogroup (Diard et al., 2010), d) it is known to have a poor survival ability in the external environment (Berthe et al., 2013). This is the first study looking at *E. coli*'s most host associated phylogroup B2 in relationship to its distribution, diversity and survival in the external environment, especially in Australian context.

The study showed that the phylogroup B2 and its predominantly human associated lineage STs 73, 95 and 131 are uncommon in Sydney and Queensland catchments (Chapter 2). Water catchments in Australia cover a vast area of land and generally are located away from urban areas or high levels of human activity. Hence, unassigned *E. coli* isolates to the predominant lineages may possibly have their origin from other sources, such as native animals, livestocks and birds or they may even be naturalised to water environment. This suggests that mere detection of *E. coli* in pristine water environments does not always mean recent human faecal contamination. If it is the case that a wide range of STs of no known host association are readily detectable in the water bodies studied, then it is also unlikely that the simple detection of *E. coli* in these water bodies reflects recent faecal contamination at all. Hence, water industries need effective markers that could differentiate *E. coli* according to its actual source of origin. The genomic comparison studies have shown that host associated phylogroup B2 can be distinguished as isolates from human source or native vertebrate source, but not naturalised to water (Chapter 3). This finding contributes to an improved understanding of the genetic diversity within phylogroup B2 and its limitations for determining the recent source of *E. coli* detected in the natural environment. Jang and colleagues, 2017 suggested that the variations in genetic regulation and expression are due to the species ecological adaptation to different hosts. This study supports the view that different ecological conditions are the driving force behind the species genetic variation as the distribution of virulence genes were more dominant in the human associated cluster (Cluster 1) and metabolic genes were more dominant in the native vertebrate associated cluster (Cluster 2) of B2 isolates. The study results are promising and it is one step closer to bridging the gap between designing the right markers for assigning *E. coli* to its source and this remains an ongoing research challenge. In summary, the study suggests the use of *eae* gene in combination with G4C genes to identify *E. coli* from human source and *appA* gene from native vertebrate source within phylogroup B2. However, this approach may need to be refined. Since phylogroup B2 is predominant in Australian humans, the *eae* and G4C genes to identify isolates from human sources seems appropriate to use. But in Australian native vertebrates, since phylogroup A, B1 and D are more frequently isolated than phylogroup B2 (Gordon and Cowling, 2003), further research on the specificity of the gene *appA* to native vertebrates is required. In general, the genetic diversity even within phylogroup B2 undermines the use of this phylogroup (broadly

defined) as a specific human FIB and adds further to the understanding of the limitations of *E. coli* as a general indicator of recent faecal contamination of water.

Of particular note, even though the human host associated phylogroup B2 isolates are rarely detected in Australian water bodies, the survival analysis suggests that these isolates can persist for a long period of time and can respond to stress by entering a VBNC state in external environments (Chapter 4). This new understanding of the persistent survival of phylogroup B2 isolates in water challenges the assumption about *E. coli's* poor environmental survivability. As a note of caution, the microcosm experiment had ideal conditions such as stable temperature and pH, and no predators or competitors. Generally, since phylogroup B2 is less often isolated from water compared to other phylogroups of *E. coli*, it leads to the hypothesis that in the natural aquatic environment phylogroup B2 are more prone to predation and competition than the other phylogroups, and some genes may play role in it. Further research is needed to test this hypothesis. Overall, the findings of this study show potential exceptions to the consideration of using *E. coli's* presence in water as an indicator of a 'recent' human faecal contamination event. It also challenges the *E. coli* detection methods used by water industries such as the COLIERT® kit as it may miss the isolates in a VBNC state and hence fail to represent the whole event in the occurrence of true human faecal contamination. Hence, the persistence of phylogroup B2 isolates in the natural environment affects the water industries in two ways: a) don't over-react to the detection of low levels of *E. coli* in natural environment as they may not represent 'recent' faecal input, b) be aware that as *E. coli* can enter into the VBNC state when stressed, the detection of low levels either in the natural environment or in drinking water systems may be explained by this mechanism.

In conclusion, the findings of this thesis show that the presence of *E. coli* in environmental waters across Australia does not necessarily reflect recent faecal contamination from humans and native vertebrates. The findings suggest that isolates from these sources can persist for long periods in the environment. Some may enter the VBNC state. At a later time, they may regrow from low levels or even levels that are not measurable with commonly used detection methods to readily detectable numbers. The implications for the water industry are that, the detection of *E. coli* in water may not represent recent faecal pollution from human or Australian native vertebrates, and more broadly recent faecal pollution in general.

Future Directions

The following are the recommendations for future research related to the findings of this thesis:

- It is clear from the survey studies performed that most of the isolates were not assigned to predominantly human associated sequence lineages. Further characterisation of these unassigned isolates to their respective ST is required to find if any particular STs within phylogroup B2 dominates survival in water.

- If there are STs in phylogroup B2 that are commonly isolated from water, designing a multiplex PCR that identifies the human dominant STs and water dominant STs of phylogroup B2 would be fruitful for easy discrimination between these isolates.

- As phylogroup A is a generalist to all vertebrates, phylogroup B1 is predominantly isolated from ectothermic vertebrates, birds and carnivorous mammals, and phylogroup D is isolated from endothermic vertebrates, a similar comparative genomics approach with the addition of water and native vertebrate isolates from phylogroups A, B1, B2 and D is necessary to provide a specific marker for isolates associated with native animals.

- A laboratory study looking at the survival of faecal *E. coli* strains in treated drinking water that has been sourced from various catchments could shed light on the likely survival of *E. coli* in distribution systems and their potential to enter the VBNC state.

- Considering the fact that phylogroup B1 is more commonly detected in water environment than phylogroup B2, a comparative microcosm experiment looking at the survival of isolates from both phylogroups simultaneously at two temperatures (10°C and 20°C) along with their predation effect would be beneficial to study if the variation in temperature and predation affects the survival dominance of phylogroup B2 in aquatic environment.

- An RNA sequencing study before an *E. coli* isolate's entry into VBNC state and after its recovery to culturability in APW, and before and after the long-term survival of isolates in FPW are needed to investigate the genes that are actively turned on / off during the process. This will give more insight on the genes that cause the variation in survival and the genes that are affected by environmental pressure.

Quote: 'The greatest risks to consumers of drinking water are pathogenic microorganisms. Protection of water sources and treatment are of paramount importance and must never be compromised'.

Australian Drinking Water Guidelines 6, 2011

References

Berthe, T. et al., 2013. Evidence for Coexistence of Distinct *Escherichia coli* Populations in Various Aquatic Environments and Their Survival in Estuary Water. *Applied and Environmental Microbiology,* 79(15), p. 4684–4693.

Blyton, M. D. et al., 2015. Genetic Structure and Antimicrobial Resistance of *Escherichia coli* and Cryptic Clades in Birds with Diverse Human Associations. *Applied and Environmental Microbiology,* 81(15), p. 5123–5133.

Brennan, F. A. et al., 2010. Characterization of Environmentally Persistent *Escherichia coli* Isolates Leached from an Irish Soil. *Applied and Environmental Microbiology,* 76(7), pp. 2175-2180.

Diard, M. et al., 2010. Pathogenicity-associated islands in extraintestinal pathogenic *Escherichia coli* are fitness elements involved in intestinal colonization. *Journal of Bacteriology,* 192(19), pp. 4885-4893.

EPA (Environmental Protection Agency) Report of Task Force on Guide Standard and Protocol for Testing Microbiological Water Purifiers. (1986).

Faust, M. A., Aotaky, A. E. & Hargadon, M. T., 1975. Effect of Physical Parameters on the In Situ Survival of *Escherichia coli* MC-6 in an Estuarine Environment. *Applied Microbiology,* 30(5), pp. 800-806.

Gordon, D. M. & Cowling, A., 2003. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology,* Volume 149, p. 3575–3586.

Gordon, D. M., O'Brien, C. L. & Pavli, P., 2015. *Escherichia coli* diversity in the lower intestinal tract of humans. *Environmental Microbiology Reports,* 7(4), pp. 642-648.

Gordon, D. M., Stern, S. E. & Collignon, P. J., 2005. Influence of the age and sex of human hosts on the distribution of *Escherichia coli* ECOR groups and virulence traits. *Microbiology,* 151(1), pp. 15-23.

Hartl, D. L. & Dykhuizen, D. E., 1984. The population genetics of *Escherichia coli*. *Annual Review of Genetics* , Volume 18, pp. 31-68.

Harwood, V. J. et al., 2014. Microbial source tracking markers for detection of fecal contamination in environmental waters: relationships between pathogens and human health outcomes. *FEMS Microbial Reviews,* Volume 38, pp. 1-40.

Ingle, D. J. et al., 2011. Biofilm Formation by and Thermal Niche and Virulence Characteristics of Escherichia spp. *Applied and Environmental Microbiology,* 77(8), pp. 2695-2700.

Jang, J. et al., 2017. Environmental *Escherichia coli*: ecology and public health implications-a review. *Journal of Applied Microbiology,* 123(3), pp. 570-581.

NHMRC, N., 2011. *Australian Drinking Water Guidelines Paper 6 National Water Quality Management Strategy.* Canberra: National Health and Medical Research Council, National Resource Management Ministerial Council, Commonwealth of Australia.

Power, M. L. et al., 2005. Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. *Environmental Microbiology,* 7(5), pp. 631-640.

Regli, S., Rose, J. B., Haas, C. N. & Gerba, C. P., 1991. Modelling risk for pathogens in drinking water. *Journal of the American Water Works Association,* 83(11), pp. 76-84.

Slanetz, L. W. & Bartley, C. H., 1957. Numbers of Enterococci in water, sewage, and feces determined by the membrane filter technique with an improved medium. *Journal of Bacteriology*, 74(5), pp. 591-595.

Walk, S. T. et al., 2007. Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. *Environmental Microbiology,* 9(9), pp. 2274-2288.

WHO (World Health Organization). Guidelines for Drinking-water Quality, Incorporating 1st and 2nd Addenda, Volume 1, Recommendations, 3rd ed.; WHO: Geneva, Switzerland (2008).

APPENDIX

A. B2 Subtyping and Doumith PCR Primers

List of primers and the size of PCR products used in the survey study of *E. coli* from Sydney and Queensland water catchments (Chapter 2) (Clermont et al., 2014; Doumith et al., 2015)

Primer	Size of amplicon (bp)	Sequence
B2 SUBTYPING		
Panel 1		
pabBgpII.f	415	5'-GAGTCACTGCCAGAAATTGCA-3'
pabBgpII.r		5'-GGCGAAAGGCTTAAAATTGCACT-3'
trpAgpIII.f	255	5'- GACGCGCTGGAATTAGGCTC-3'
trpAgpIII.r		5'-ATCGGCAACCAGCACCGAAT-3'
dinBgpVI.f	652	5'-CAGCGGTGGAGATGCGCGAT-3'
dinBgpVI.r		5'-TCGTCAATGCCCTGACTACA-3'
icdgpVII.f	810	5'-GCGGTATTCGCTCTCTGAAT-3'
icdgpVII.r		5'-CAATTAAATCAGCCGCTTCG-3'
aesgpIX.f	160	5'-CCTGGCCTGCAACGGGAG-3'
aesgpIX.r		5'-TCTGGCTGCGGATAAAAGAG-3'
chuAgene.1	1013	5'-CGATACGGTCGATGCAAAAG-3'
chuAgene.2		5'-TTGGACAACATCAGGTCATC-3'
Panel 2		
putPgpI.f	373	5'-GGTATCGCTTACTTTAACGG-3'
putPgpI.r		5'-ACCACCGGACCAAACGCC-3'
trpAgpIV.f	261	5'-TGCCAGTGGAAGAGTCCGCT-3'
trpAgpIV.r		5'-CCGGGGCGGAAATACCAAAG-3'
polBgpV.f	530	5'-GCCGTTTCGCCGAAGATAAA-3'
polBgpV.r		5'-TAATGATCTTCAGCGCCTGT-3'
aesgpX.f	713	5'-GACCGTTGTGAATACTCTTCA-3'
aesgpX.r		5'-TATAACAGGGCGGCACATTT-3'
chuAgene.1	1013	5'-CGATACGGTCGATGCAAAAG-3'
chuAgene.2		5'-TTGGACAACATCAGGTCATC-3'
DOUMITH		
ST73_for	490	5'-TGGTTTTACCATTTTGTCGGA-3'
ST73_rev		5'-GGAAATCGTTGATGTTGGCT-3'
ST131_for	310	5'-GACTGCATTTCGTCGCCATA-3'
ST131_rev		5'-CCGGCGGCATCATAATGAAA-3'
ST95_for	200	5'-ACTAATCAGGATGGCGAGAC-3'
ST95_rev		5'-ATCACGCCCATTAATCCAGT-3'
ST69_for	104	5'-ATCTGGAGGCAACAAGCATA-3'
ST69_rev		5'-AGAGAAAGGGCGTTCAGAAT-3'

References

Clermont, O. et al., 2014. Development of an allele-specific PCR for *Escherichia coli* B2 sub-typing, a rapid and easy to perform substitute of multilocus sequence typing. Journal of Microbiological Methods, Volume 101, pp. 24-27

Doumith, M. et al., 2015. Rapid Identification of Major *Escherichia coli* Sequence Types Causing Urinary Tract and Bloodstream Infections. Journal of Clinical Microbiology, 53(1), pp. 160-166

B. Percentage of each genes in Cluster 1 and Cluster 2

Percentage of each genes represented in strains from Cluster 1 and Cluster 2 in comparative genomic studies (Chapter 3).

Gene	Cluster 1 (%)	Cluster 2 (%)
aam	38%	6%
aas_2	13%	18%
abgA_1	22%	6%
abgB	22%	6%
abgR	22%	6%
abgT	22%	6%
acpP_3	13%	18%
acpT_2	38%	6%
adhE_2	34%	88%
aes_2	78%	82%
agaC_2	31%	6%
agaR_3	31%	71%
agaR_4	9%	15%
agaS_2	9%	15%
aldA	100%	85%
alkA	72%	74%
alpA	34%	53%
alpA_1	19%	35%
alsC	91%	82%
ampC_2	38%	6%
appA	38%	100%
appY	6%	18%
araE	81%	91%
ariR	72%	76%
ascB	72%	56%
ascF	69%	56%
ascG	72%	56%
aslA	56%	68%
besA	28%	18%
bfpA	34%	47%
bfpB	9%	15%
bglA_1	72%	88%
bglG_1	28%	71%
bglH_2	72%	100%

Gene	Cluster 1 (%)	Cluster 2 (%)
btuB_2	72%	85%
btuF_3	75%	44%
cadC	94%	76%
cadC_2	19%	29%
cai	9%	24%
caiA_2	38%	6%
caiT_2	53%	82%
cas1	41%	56%
cas2	25%	0%
cas3	41%	56%
cas6f	41%	56%
casA	25%	0%
casC	25%	0%
casD	25%	0%
casE	25%	0%
cba	6%	21%
cbeA	13%	3%
cbeA_1	44%	21%
cbeA_2	25%	26%
cbtA	72%	9%
cbtA_1	9%	29%
cbtA_2	13%	15%
cca_1	59%	41%
ccdA_2	41%	12%
ccdB_2	41%	12%
cdiA	25%	21%
cea	6%	18%
cfaB	13%	32%
cfaE	13%	32%
chbA_2	88%	91%
chbB_2	84%	91%
chbC_2	84%	91%
cheA	88%	85%
cia	22%	9%
cirA_4	9%	15%
clcB	75%	91%
clpB_3	31%	62%
clpP_2	31%	21%
cma	9%	24%
cmi	9%	24%
cotSA	41%	0%
cra_1	41%	41%
cspB	9%	21%
cspF	19%	18%
cspH	31%	6%
cspI	6%	21%
csy1	41%	56%
csy2	41%	56%
csy2_1	25%	6%

Gene	Cluster 1 (%)	Cluster 2 (%)
csy3	41%	56%
cvaA	13%	21%
cydA	63%	68%
cydC	47%	62%
cydC_2	28%	29%
cytR_1	63%	12%
cytR_2	34%	32%
dam_1	22%	41%
dapA_2	31%	6%
dcuC	75%	82%
ddrA	34%	88%
deoR_2	22%	24%
dgoA_2	69%	76%
dgoK_2	69%	76%
dhaK_2	53%	82%
dhaL_2	53%	82%
dhaR_2	50%	79%
dicC	16%	12%
dinD	50%	74%
dinI_1	22%	44%
dinI_2	22%	24%
dinI_3	22%	18%
dkgA_3	84%	85%
dltC	38%	6%
dnaB_1	16%	15%
dnaQ_1	13%	15%
dnaT_2	6%	12%
dpiB	97%	76%
dppA	78%	65%
dsdC	44%	50%
dxs_2	6%	29%
efeB	88%	62%
elfC	91%	62%
elfG	34%	21%
elmGT	28%	18%
emrB_1	78%	94%
emrE_1	13%	18%
emrK_1	78%	94%
entD	94%	71%
entE_4	38%	6%
envR	88%	91%
essD	6%	38%
essD_1	28%	0%
essQ	13%	24%
etk	69%	0%
etp	69%	0%
eutE_2	34%	88%
eutK_2	34%	88%
eutL_2	34%	88%

Gene	Cluster 1 (%)	Cluster 2 (%)
eutM_2	34%	88%
eutN_2	34%	88%
eutS_2	34%	88%
fabD_2	38%	6%
fabG_3	31%	71%
fabZ_2	6%	18%
fadD_1	78%	88%
fadD_2	94%	91%
fadH_3	9%	3%
fadK_2	38%	6%
fadL	72%	59%
fdoH	66%	65%
fdtC	13%	9%
fecD_5	72%	44%
fepA_2	25%	18%
fepC_3	66%	44%
fes_2	28%	18%
fhlA	75%	82%
fhlA_2	19%	29%
fhuA_2	53%	38%
fhuA_3	75%	44%
fhuC_3	75%	74%
fimA_2	28%	18%
fimB_1	75%	97%
fimD_1	25%	53%
fimF_2	9%	15%
fimI_2	88%	65%
fimI_3	34%	35%
fimI_4	16%	29%
fimZ_4	75%	85%
fimZ_5	28%	15%
finO	63%	21%
flgB_2	19%	29%
flgC_2	16%	29%
flgD_2	19%	29%
flgE_2	19%	29%
flgF_2	19%	29%
flgG_2	19%	29%
flgH_2	19%	29%
flgI_2	19%	29%
flgJ_2	19%	29%
flgK_2	19%	29%
flgL_2	19%	29%
flhB_2	19%	29%
fliA_2	19%	32%
fliC	13%	35%
fliC_2	19%	29%
fliD	81%	59%
fliD_2	19%	32%

Gene	Cluster 1 (%)	Cluster 2 (%)
fliE_2	19%	29%
fliF_2	19%	29%
fliG_2	19%	29%
fliH_2	19%	29%
fliI_2	19%	29%
fliN_2	19%	29%
fliP_2	19%	29%
fliQ_2	19%	29%
fliR_2	19%	29%
fliS_2	19%	32%
flu	22%	9%
flu_2	13%	12%
folK	91%	91%
fucA_2	41%	3%
fucK_2	63%	53%
fucP_2	44%	29%
gadB	75%	79%
gadB_1	59%	65%
galE_2	19%	21%
gapA_2	47%	88%
garD_2	9%	12%
garD_4	9%	12%
gatA_2	31%	71%
gatA_3	31%	71%
gatB_2	63%	50%
gatC_1	69%	68%
gatC_2	63%	53%
gatC_3	31%	71%
gcl_2	50%	76%
gfcA	69%	0%
gfcB	69%	0%
gfcD	69%	0%
gfcE	25%	94%
ghrB_1	81%	91%
gldA_1	53%	82%
glnL_7	25%	41%
glpF_2	34%	85%
glpK_1	6%	29%
glpT_3	25%	38%
glxK_2	53%	82%
glxR_2	53%	79%
gnsB	6%	24%
gntP	69%	32%
gntR_2	44%	29%
gpFI	31%	44%
gpFI_1	25%	44%
group_1	25%	32%
group_10	22%	6%
group_1010	3%	15%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_1011	53%	21%
group_1012	53%	18%
group_1013	91%	82%
group_1014	53%	59%
group_10145	9%	15%
group_10156	9%	15%
group_10157	9%	15%
group_10158	9%	12%
group_10159	9%	15%
group_10166	6%	15%
group_1018	44%	0%
group_10186	3%	12%
group_10188	9%	15%
group_1019	22%	29%
group_10190	9%	12%
group_10191	9%	12%
group_10192	9%	12%
group_1020	19%	21%
group_1021	25%	24%
group_10217	3%	12%
group_10218	3%	12%
group_10219	3%	12%
group_10221	3%	12%
group_10222	3%	12%
group_10223	3%	12%
group_10224	3%	12%
group_10225	3%	12%
group_10226	3%	12%
group_10227	3%	12%
group_10228	3%	9%
group_10232	3%	12%
group_10235	3%	12%
group_10236	3%	12%
group_1024	66%	32%
group_1026	25%	18%
group_10277	3%	18%
group_10278	3%	18%
group_10279	3%	18%
group_1028	78%	97%
group_10281	9%	12%
group_10282	9%	12%
group_10283	9%	12%
group_1030	16%	21%
group_10303	3%	18%
group_1032	28%	9%
group_1033	9%	6%
group_1037	22%	18%
group_1043	25%	18%
group_1044	78%	91%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_1047	22%	3%
group_1048	22%	3%
group_1049	13%	15%
group_1053	13%	26%
group_10622	16%	15%
group_10636	22%	3%
group_10637	22%	3%
group_10724	22%	18%
group_10743	3%	3%
group_1078	31%	53%
group_10839	3%	6%
group_10840	3%	6%
group_1093	78%	85%
group_1095	34%	15%
group_1098	22%	3%
group_1104	19%	6%
group_11043	13%	0%
group_1108	38%	26%
group_11098	13%	0%
group_11099	13%	0%
group_1110	28%	47%
group_1111	25%	21%
group_1114	66%	26%
group_11145	13%	0%
group_11146	13%	0%
group_11147	6%	0%
group_1115	53%	24%
group_11157	13%	0%
group_11158	13%	0%
group_1118	16%	9%
group_1120	9%	9%
group_1129	16%	9%
group_1132	31%	18%
group_1134	16%	12%
group_11346	0%	3%
group_11351	0%	6%
group_11352	0%	6%
group_11396	3%	6%
group_11407	9%	6%
group_1145	31%	26%
group_1146	6%	15%
group_1148	50%	21%
group_1150	6%	24%
group_116	13%	32%
group_1161	22%	6%
group_1162	31%	18%
group_1167	19%	6%
group_1168	31%	29%
group_1170	31%	50%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_1171	31%	53%
group_1172	31%	41%
group_1184	16%	12%
group_1189	25%	9%
group_1195	13%	18%
group_1198	25%	26%
group_1200	28%	29%
group_1203	31%	18%
group_1215	0%	9%
group_1218	31%	21%
group_1221	9%	24%
group_1222	28%	18%
group_1224	84%	85%
group_1226	38%	32%
group_123	3%	29%
group_1232	25%	38%
group_1240	31%	6%
group_1241	9%	18%
group_1244	6%	15%
group_1247	53%	9%
group_1250	16%	24%
group_1255	66%	32%
group_1257	19%	15%
group_1260	53%	21%
group_1268	28%	74%
group_1273	59%	74%
group_1280	6%	18%
group_1281	34%	15%
group_1282	44%	21%
group_1283	9%	9%
group_1287	28%	26%
group_129	16%	9%
group_1295	38%	76%
group_13	13%	9%
group_1307	28%	50%
group_1322	19%	29%
group_1333	78%	97%
group_1338	28%	15%
group_1339	25%	44%
group_1340	22%	44%
group_1346	13%	12%
group_1349	16%	44%
group_136	22%	6%
group_1361	3%	15%
group_1368	19%	29%
group_1375	34%	21%
group_1385	59%	24%
group_1386	59%	56%
group_1388	25%	41%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_1392	50%	32%
group_1393	31%	18%
group_1397	16%	18%
group_1398	38%	3%
group_1399	63%	32%
group_1401	25%	18%
group_1406	13%	6%
group_1408	16%	12%
group_1415	3%	9%
group_14263	38%	6%
group_143	6%	21%
group_1434	63%	3%
group_1439	56%	6%
group_1448	19%	0%
group_1451	13%	21%
group_1453	13%	6%
group_1460	25%	12%
group_1461	16%	6%
group_1466	22%	18%
group_1468	38%	6%
group_1475	6%	18%
group_1476	9%	24%
group_1478	3%	18%
group_1479	16%	15%
group_1484	9%	15%
group_1487	25%	9%
group_1492	25%	0%
group_1494	19%	18%
group_1498	13%	26%
group_1499	75%	68%
group_15	9%	15%
group_1511	25%	94%
group_1514	28%	6%
group_1515	56%	71%
group_1516	31%	47%
group_1517	9%	15%
group_1523	19%	6%
group_1524	9%	18%
group_1525	6%	18%
group_1529	6%	26%
group_1531	3%	21%
group_1536	41%	59%
group_1537	16%	21%
group_1542	22%	38%
group_1546	6%	12%
group_1547	0%	6%
group_1561	28%	21%
group_1567	22%	3%
group_1569	41%	3%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_157	28%	6%
group_1571	22%	3%
group_1572	38%	32%
group_1575	22%	9%
group_1583	34%	21%
group_1586	66%	26%
group_1588	31%	35%
group_1589	25%	41%
group_1590	28%	44%
group_1591	50%	41%
group_160	22%	6%
group_1601	28%	18%
group_1602	31%	18%
group_16026	66%	56%
group_1604	19%	9%
group_1609	6%	24%
group_161	13%	12%
group_1611	88%	79%
group_1612	38%	21%
group_1613	19%	9%
group_1615	66%	18%
group_1616	25%	35%
group_1621	25%	3%
group_1627	25%	3%
group_1634	31%	12%
group_1645	16%	6%
group_1648	16%	3%
group_1655	41%	26%
group_1656	19%	15%
group_16567	3%	47%
group_16570	0%	12%
group_16572	9%	6%
group_16574	0%	6%
group_16579	25%	35%
group_1660	16%	9%
group_1665	13%	24%
group_1674	41%	59%
group_1675	16%	35%
group_1678	16%	26%
group_1682	22%	24%
group_1685	38%	74%
group_1686	19%	12%
group_1688	6%	3%
group_1690	31%	26%
group_1693	13%	9%
group_1696	50%	18%
group_1697	44%	91%
group_1699	25%	12%
group_17	16%	15%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_1700	31%	12%
group_1703	16%	12%
group_1704	9%	18%
group_1705	47%	79%
group_1710	25%	18%
group_1715	13%	32%
group_1717	25%	76%
group_1718	0%	26%
group_1723	31%	41%
group_1724	41%	74%
group_1725	9%	15%
group_1726	9%	15%
group_1729	6%	26%
group_173	13%	15%
group_1735	3%	18%
group_1736	13%	18%
group_1737	19%	44%
group_1744	3%	21%
group_1748	9%	9%
group_175	6%	15%
group_1751	13%	15%
group_1753	3%	12%
group_1755	41%	0%
group_1769	19%	9%
group_1770	3%	12%
group_1771	6%	18%
group_1773	3%	9%
group_1775	63%	26%
group_1779	13%	9%
group_1780	88%	79%
group_1781	19%	21%
group_1783	25%	44%
group_1784	25%	41%
group_1795	59%	41%
group_1796	6%	21%
group_1797	22%	24%
group_18072	6%	18%
group_1808	59%	18%
group_1811	0%	9%
group_1817	22%	12%
group_18224	84%	65%
group_18225	3%	21%
group_1823	41%	3%
group_1826	28%	0%
group_1827	72%	94%
group_183	9%	18%
group_1839	41%	44%
group_1845	13%	12%
group_1848	6%	18%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_1852	63%	9%
group_1854	13%	12%
group_1856	63%	59%
group_1862	59%	26%
group_1864	41%	24%
group_1866	13%	12%
group_1867	28%	44%
group_1868	31%	44%
group_1869	44%	41%
group_1870	13%	9%
group_1874	34%	0%
group_1877	47%	0%
group_1881	56%	21%
group_1882	53%	15%
group_1884	19%	32%
group_1893	6%	15%
group_1895	19%	18%
group_1896	13%	21%
group_1902	28%	15%
group_1903	25%	12%
group_1906	84%	74%
group_1907	16%	29%
group_1915	13%	15%
group_1919	38%	3%
group_1924	59%	6%
group_1926	22%	3%
group_1927	66%	9%
group_1935	19%	6%
group_1940	22%	18%
group_1942	6%	6%
group_1948	31%	18%
group_1958	31%	6%
group_1960	19%	15%
group_1961	16%	12%
group_1962	38%	56%
group_1969	6%	15%
group_1974	16%	29%
group_1976	22%	32%
group_1989	19%	12%
group_1990	25%	44%
group_1991	6%	24%
group_1992	25%	29%
group_1993	19%	56%
group_2000	22%	24%
group_2001	22%	6%
group_2002	16%	9%
group_2003	16%	9%
group_2009	34%	53%
group_2010	6%	21%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_2011	47%	29%
group_2012	13%	18%
group_2013	19%	6%
group_2014	31%	50%
group_2015	16%	18%
group_2017	13%	12%
group_2025	47%	71%
group_2033	53%	35%
group_2038	3%	3%
group_2043	16%	47%
group_2045	31%	41%
group_206	9%	15%
group_207	19%	26%
group_2073	9%	12%
group_2074	6%	9%
group_2080	9%	21%
group_2081	6%	18%
group_2083	0%	12%
group_2086	3%	6%
group_2092	16%	12%
group_2095	28%	26%
group_2096	3%	21%
group_2098	19%	29%
group_2103	22%	32%
group_2109	9%	15%
group_211	13%	6%
group_2112	3%	29%
group_2118	3%	12%
group_214	6%	6%
group_2142	59%	38%
group_2149	19%	15%
group_2167	22%	24%
group_2171	25%	32%
group_2186	6%	18%
group_2189	38%	44%
group_219	9%	12%
group_2194	31%	6%
group_2198	38%	65%
group_2215	88%	85%
group_2216	13%	15%
group_2224	22%	21%
group_2227	19%	9%
group_2235	28%	9%
group_2239	22%	9%
group_2244	19%	29%
group_2254	72%	41%
group_2262	38%	0%
group_227	13%	15%
group_2271	25%	15%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_2280	78%	97%
group_2288	69%	21%
group_2291	16%	21%
group_2295	53%	32%
group_2310	9%	21%
group_2314	22%	32%
group_2321	22%	15%
group_2322	66%	26%
group_2334	31%	9%
group_2339	13%	12%
group_2340	31%	0%
group_2343	59%	9%
group_2350	28%	18%
group_2352	31%	21%
group_2353	63%	21%
group_2354	22%	0%
group_2360	22%	9%
group_2361	47%	21%
group_2373	38%	24%
group_2374	41%	15%
group_2378	9%	15%
group_2382	69%	68%
group_2387	34%	18%
group_2390	31%	18%
group_2396	91%	82%
group_2397	13%	18%
group_2398	16%	3%
group_2404	25%	0%
group_2410	19%	12%
group_2412	63%	3%
group_2415	19%	12%
group_2421	19%	9%
group_2426	22%	32%
group_243	22%	15%
group_2431	88%	82%
group_2433	16%	18%
group_2443	53%	82%
group_2446	53%	18%
group_2447	6%	24%
group_2453	25%	3%
group_2459	22%	15%
group_246	16%	24%
group_2460	16%	6%
group_2462	16%	3%
group_2463	16%	15%
group_2464	19%	12%
group_2469	22%	15%
group_2472	31%	3%
group_2473	25%	9%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_2480	22%	3%
group_2481	13%	21%
group_2483	9%	15%
group_2486	16%	6%
group_2487	41%	91%
group_2490	31%	12%
group_2493	28%	3%
group_2494	19%	6%
group_2495	31%	21%
group_2514	19%	12%
group_2523	13%	12%
group_2528	16%	15%
group_2532	13%	15%
group_2533	13%	12%
group_2534	19%	32%
group_2536	0%	24%
group_2538	31%	21%
group_2544	16%	12%
group_256	19%	21%
group_2563	19%	29%
group_2576	28%	26%
group_2586	19%	32%
group_2588	25%	35%
group_2592	9%	18%
group_2593	9%	18%
group_2596	38%	41%
group_2605	3%	12%
group_2607	3%	12%
group_2639	41%	32%
group_2642	16%	18%
group_2644	19%	9%
group_2659	6%	24%
group_2661	3%	26%
group_2675	25%	44%
group_2676	25%	44%
group_2683	50%	12%
group_2688	13%	38%
group_2698	63%	56%
group_2707	13%	18%
group_2724	16%	12%
group_2728	28%	21%
group_2731	13%	12%
group_2733	28%	9%
group_2747	6%	21%
group_2750	3%	18%
group_2759	31%	0%
group_2782	6%	18%
group_2786	13%	18%
group_2794	3%	15%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_2804	22%	6%
group_2814	66%	79%
group_2815	34%	24%
group_2818	22%	38%
group_2819	53%	15%
group_284	34%	3%
group_2841	13%	18%
group_2844	44%	32%
group_285	19%	21%
group_2851	34%	3%
group_2853	41%	47%
group_2860	19%	9%
group_2871	6%	18%
group_288	22%	15%
group_2886	13%	18%
group_2887	9%	18%
group_2888	9%	26%
group_2889	25%	38%
group_2892	38%	15%
group_2893	66%	26%
group_2895	66%	26%
group_2901	50%	12%
group_2905	47%	18%
group_2908	16%	9%
group_2909	16%	32%
group_291	34%	21%
group_2910	31%	44%
group_2912	19%	26%
group_2916	25%	26%
group_292	22%	18%
group_2927	31%	21%
group_2929	19%	3%
group_293	13%	12%
group_2931	53%	65%
group_2932	63%	32%
group_2934	19%	9%
group_2944	31%	12%
group_2947	28%	15%
group_296	6%	9%
group_2978	50%	6%
group_298	50%	21%
group_2980	31%	3%
group_2982	59%	6%
group_2995	22%	24%
group_2997	28%	44%
group_2999	19%	18%
group_3002	31%	18%
group_3005	34%	18%
group_3009	31%	9%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_3015	28%	68%
group_3017	25%	6%
group_3027	31%	0%
group_3028	31%	0%
group_3029	31%	35%
group_3037	19%	3%
group_3038	6%	24%
group_3041	19%	6%
group_3042	22%	9%
group_3052	50%	6%
group_3059	6%	18%
group_306	9%	9%
group_3060	47%	24%
group_3073	22%	41%
group_3075	3%	32%
group_3078	59%	94%
group_308	28%	26%
group_3081	22%	9%
group_3082	25%	3%
group_3083	22%	9%
group_3084	31%	82%
group_3086	16%	35%
group_3088	22%	6%
group_3091	22%	6%
group_3093	41%	91%
group_3094	41%	91%
group_3096	25%	12%
group_3097	44%	56%
group_3101	19%	15%
group_3104	13%	12%
group_3107	16%	32%
group_3113	19%	18%
group_3114	25%	3%
group_3115	22%	3%
group_3118	38%	38%
group_3119	44%	59%
group_3120	47%	62%
group_3121	16%	9%
group_3122	22%	6%
group_3123	16%	15%
group_3127	22%	26%
group_3128	19%	24%
group_3131	19%	12%
group_3136	22%	3%
group_3139	3%	18%
group_3142	28%	3%
group_3143	22%	18%
group_3144	38%	6%
group_3146	13%	12%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_3147	13%	12%
group_3149	13%	12%
group_315	22%	12%
group_3151	13%	12%
group_3153	13%	12%
group_3155	41%	41%
group_3172	6%	15%
group_3182	6%	6%
group_3185	9%	65%
group_3191	16%	35%
group_3192	13%	24%
group_3194	6%	24%
group_3200	28%	26%
group_3209	0%	12%
group_3225	0%	32%
group_3226	19%	29%
group_3239	22%	18%
group_3255	28%	0%
group_3257	31%	29%
group_3274	9%	15%
group_3276	13%	6%
group_328	56%	24%
group_3281	16%	15%
group_329	31%	26%
group_3294	59%	12%
group_3300	3%	12%
group_3301	6%	18%
group_3307	16%	18%
group_331	22%	9%
group_3313	66%	47%
group_3316	6%	6%
group_3317	6%	6%
group_332	28%	0%
group_3321	3%	6%
group_3322	3%	6%
group_3327	28%	9%
group_3329	3%	6%
group_3330	3%	15%
group_3352	13%	9%
group_3397	81%	88%
group_3405	28%	18%
group_3407	9%	12%
group_3411	13%	9%
group_3428	16%	26%
group_3429	25%	41%
group_3431	19%	38%
group_3432	25%	44%
group_3439	53%	44%
group_3442	66%	38%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_3444	53%	38%
group_345	22%	15%
group_3455	28%	91%
group_3456	72%	12%
group_3460	56%	6%
group_3476	44%	29%
group_3480	19%	18%
group_3502	13%	18%
group_3538	72%	62%
group_3541	63%	9%
group_3549	9%	12%
group_3551	13%	12%
group_3567	34%	32%
group_3583	9%	18%
group_3587	66%	85%
group_359	6%	38%
group_3590	16%	12%
group_3591	25%	21%
group_3592	28%	21%
group_3593	28%	21%
group_361	25%	21%
group_3615	34%	24%
group_363	72%	71%
group_364	13%	9%
group_3641	3%	0%
group_3657	13%	18%
group_3660	13%	18%
group_3662	28%	24%
group_3666	66%	32%
group_3671	53%	0%
group_3678	53%	15%
group_3705	16%	41%
group_3706	38%	26%
group_3710	9%	32%
group_3720	41%	35%
group_3737	78%	97%
group_3747	41%	38%
group_3779	22%	12%
group_3815	38%	35%
group_382	13%	9%
group_3821	69%	79%
group_3826	9%	18%
group_3827	9%	21%
group_3829	9%	21%
group_3832	9%	21%
group_3837	9%	24%
group_3839	25%	35%
group_3842	25%	6%
group_3843	63%	21%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_3844	66%	26%
group_3845	13%	6%
group_385	9%	18%
group_3852	31%	21%
group_3859	69%	18%
group_3861	28%	6%
group_3862	28%	44%
group_3864	31%	44%
group_3866	31%	38%
group_3868	31%	38%
group_3870	31%	44%
group_3872	25%	44%
group_3877	34%	38%
group_3888	59%	9%
group_3892	6%	15%
group_3895	31%	18%
group_3898	28%	18%
group_3900	31%	21%
group_3905	63%	56%
group_3906	16%	18%
group_3907	6%	15%
group_3910	22%	9%
group_3913	63%	32%
group_3918	16%	15%
group_3920	19%	9%
group_3921	25%	15%
group_3929	34%	18%
group_3930	28%	15%
group_3932	31%	18%
group_3933	41%	0%
group_3935	44%	0%
group_3936	13%	9%
group_3938	31%	0%
group_3939	34%	12%
group_3940	38%	3%
group_3941	34%	18%
group_3943	38%	0%
group_3946	84%	79%
group_3951	19%	3%
group_3957	16%	15%
group_3960	28%	21%
group_3961	19%	9%
group_3962	22%	18%
group_3978	59%	97%
group_3982	97%	74%
group_3984	41%	15%
group_3985	28%	0%
group_3986	22%	6%
group_400	9%	18%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_4000	9%	15%
group_4010	19%	26%
group_4011	59%	3%
group_4019	13%	9%
group_4025	34%	12%
group_4030	78%	94%
group_4036	53%	9%
group_404	3%	15%
group_4043	6%	9%
group_4059	16%	3%
group_4060	16%	3%
group_4063	34%	29%
group_4071	69%	47%
group_4075	59%	41%
group_4077	38%	3%
group_4081	25%	0%
group_4087	78%	91%
group_4125	22%	18%
group_4127	22%	6%
group_4130	13%	15%
group_4134	22%	50%
group_4147	9%	12%
group_4148	25%	3%
group_4149	22%	6%
group_4150	13%	56%
group_4152	31%	88%
group_4159	41%	91%
group_4161	31%	62%
group_4162	31%	62%
group_4164	31%	71%
group_4166	31%	56%
group_4167	22%	44%
group_4171	31%	65%
group_4174	41%	91%
group_4176	19%	9%
group_4183	22%	6%
group_4186	22%	3%
group_4191	25%	6%
group_4194	19%	18%
group_4207	22%	26%
group_4216	22%	3%
group_4218	13%	21%
group_4221	38%	12%
group_4224	13%	12%
group_4225	13%	12%
group_4227	13%	12%
group_4252	6%	18%
group_4257	34%	88%
group_4258	34%	88%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_4268	13%	15%
group_4269	31%	50%
group_4271	16%	6%
group_4278	13%	24%
group_4279	25%	38%
group_4283	0%	24%
group_4287	9%	12%
group_4293	22%	0%
group_4296	0%	26%
group_431	25%	26%
group_4311	31%	71%
group_4326	13%	18%
group_4341	16%	9%
group_4343	16%	6%
group_4344	0%	24%
group_4345	22%	15%
group_4350	19%	29%
group_4351	19%	29%
group_4353	16%	29%
group_4354	19%	32%
group_4355	19%	32%
group_4361	0%	35%
group_4366	19%	24%
group_4367	25%	44%
group_4369	22%	3%
group_4372	31%	0%
group_4375	25%	15%
group_4385	28%	38%
group_4388	13%	12%
group_4389	31%	26%
group_439	53%	12%
group_4397	0%	35%
group_4401	25%	21%
group_4406	38%	32%
group_441	9%	12%
group_4424	9%	18%
group_4425	9%	15%
group_4426	9%	18%
group_4427	9%	18%
group_4429	9%	18%
group_4432	9%	15%
group_4435	9%	12%
group_4446	31%	59%
group_445	22%	3%
group_4452	6%	24%
group_4456	0%	29%
group_4459	16%	35%
group_4460	13%	9%
group_4464	9%	21%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_4470	19%	6%
group_4478	9%	15%
group_4481	6%	15%
group_4484	9%	12%
group_4485	3%	12%
group_4486	3%	12%
group_4487	3%	12%
group_4488	3%	12%
group_4489	3%	12%
group_4490	3%	12%
group_45	13%	47%
group_4513	6%	3%
group_4526	3%	6%
group_4530	0%	6%
group_4558	13%	6%
group_4569	13%	26%
group_472	28%	3%
group_4767	22%	44%
group_4769	25%	44%
group_477	16%	18%
group_4771	25%	44%
group_4772	25%	44%
group_4773	25%	44%
group_4796	69%	0%
group_4833	66%	53%
group_484	22%	24%
group_4852	47%	62%
group_487	22%	26%
group_488	41%	0%
group_493	22%	9%
group_494	6%	24%
group_4945	19%	12%
group_4947	19%	18%
group_4950	22%	21%
group_4953	25%	15%
group_4954	28%	21%
group_4955	25%	15%
group_4956	25%	15%
group_4985	25%	6%
group_4993	16%	6%
group_5019	44%	56%
group_504	13%	12%
group_5049	38%	29%
group_5076	13%	18%
group_5079	13%	18%
group_5081	13%	18%
group_5083	13%	18%
group_5085	25%	0%
group_5105	53%	29%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_514	31%	44%
group_516	25%	38%
group_5174	28%	44%
group_522	25%	12%
group_5221	78%	91%
group_5222	63%	6%
group_523	6%	21%
group_525	13%	12%
group_527	9%	9%
group_528	28%	21%
group_5282	84%	85%
group_5288	9%	18%
group_5289	9%	18%
group_5290	9%	21%
group_5293	9%	24%
group_5294	9%	24%
group_5295	19%	18%
group_5297	16%	15%
group_5298	9%	21%
group_5301	38%	21%
group_5302	41%	21%
group_5303	66%	26%
group_5304	3%	6%
group_5305	28%	18%
group_5319	31%	44%
group_5321	31%	44%
group_5322	13%	15%
group_5346	66%	35%
group_5359	25%	12%
group_5360	31%	18%
group_5362	19%	6%
group_5365	28%	0%
group_5367	22%	6%
group_5368	22%	6%
group_537	56%	76%
group_5384	13%	15%
group_5389	34%	6%
group_5392	34%	18%
group_5394	34%	18%
group_5396	25%	3%
group_5397	13%	15%
group_5398	41%	0%
group_5399	41%	9%
group_5401	44%	0%
group_5409	22%	3%
group_5410	31%	18%
group_5415	38%	0%
group_5416	31%	3%
group_5434	22%	9%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_5435	28%	21%
group_5436	28%	21%
group_5465	9%	15%
group_5467	9%	15%
group_5469	9%	15%
group_5483	66%	9%
group_5495	56%	26%
group_550	28%	12%
group_5505	31%	44%
group_5507	31%	41%
group_5515	13%	9%
group_5516	13%	9%
group_5520	31%	18%
group_5535	31%	9%
group_5537	9%	15%
group_554	19%	9%
group_5549	16%	12%
group_5565	16%	3%
group_5568	13%	3%
group_5569	28%	0%
group_5571	19%	3%
group_5574	25%	6%
group_5578	38%	12%
group_558	6%	18%
group_5580	38%	0%
group_5584	59%	32%
group_5585	59%	32%
group_5588	31%	35%
group_5590	69%	44%
group_5597	41%	32%
group_5609	13%	32%
group_5610	25%	56%
group_5619	13%	21%
group_5620	22%	18%
group_5629	6%	18%
group_5630	6%	18%
group_5649	47%	59%
group_5670	25%	44%
group_5671	34%	9%
group_5673	25%	6%
group_5674	22%	9%
group_5675	19%	9%
group_5676	19%	9%
group_5677	19%	9%
group_5678	19%	15%
group_5683	75%	56%
group_5684	25%	6%
group_5685	22%	15%
group_5688	31%	65%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_5691	22%	9%
group_5692	38%	6%
group_5697	25%	0%
group_5698	25%	3%
group_5701	22%	6%
group_5702	22%	3%
group_5708	16%	6%
group_5710	22%	3%
group_5713	34%	6%
group_5715	38%	6%
group_5718	50%	44%
group_5719	25%	91%
group_5720	19%	12%
group_5721	13%	9%
group_5722	6%	15%
group_5723	19%	3%
group_5729	44%	91%
group_5730	31%	38%
group_5731	31%	65%
group_5732	31%	65%
group_5733	31%	65%
group_5734	31%	65%
group_5735	31%	65%
group_5736	31%	65%
group_5737	31%	65%
group_5738	31%	65%
group_5743	16%	24%
group_5746	28%	18%
group_5749	25%	41%
group_5750	22%	50%
group_5756	9%	12%
group_5761	22%	12%
group_5769	31%	6%
group_577	16%	18%
group_5776	16%	9%
group_5798	22%	3%
group_5821	31%	29%
group_5827	41%	3%
group_5828	38%	3%
group_5829	25%	21%
group_5833	16%	9%
group_5837	31%	18%
group_584	3%	12%
group_5841	25%	6%
group_5843	3%	24%
group_5845	13%	12%
group_5847	13%	12%
group_5850	13%	12%
group_5852	13%	12%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_5853	13%	12%
group_5854	13%	12%
group_5856	13%	12%
group_5857	13%	12%
group_586	25%	18%
group_5860	16%	12%
group_587	28%	18%
group_588	34%	12%
group_5900	34%	88%
group_5904	34%	88%
group_5908	0%	21%
group_5909	6%	56%
group_591	44%	91%
group_5916	25%	35%
group_5927	34%	88%
group_5935	6%	35%
group_594	41%	91%
group_5942	3%	21%
group_5944	3%	26%
group_5947	0%	21%
group_5967	63%	53%
group_5972	34%	79%
group_6023	3%	50%
group_6037	9%	0%
group_6049	13%	9%
group_6054	0%	24%
group_6065	19%	29%
group_6093	16%	9%
group_6105	22%	41%
group_6106	13%	24%
group_6109	9%	26%
group_6119	34%	35%
group_6120	3%	26%
group_6134	28%	18%
group_6136	22%	15%
group_6141	22%	18%
group_6142	22%	12%
group_6144	28%	15%
group_6147	19%	9%
group_615	3%	21%
group_6170	28%	26%
group_6174	25%	26%
group_6175	31%	35%
group_618	22%	15%
group_6184	19%	29%
group_6189	19%	24%
group_6191	13%	12%
group_6194	3%	44%
group_62	3%	9%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_6200	22%	29%
group_6202	25%	29%
group_6203	25%	26%
group_6205	0%	24%
group_621	13%	26%
group_6217	9%	18%
group_6219	9%	18%
group_6220	9%	18%
group_6221	9%	18%
group_6223	9%	18%
group_6224	6%	15%
group_6232	6%	6%
group_6240	6%	6%
group_626	56%	12%
group_6272	0%	35%
group_6277	6%	24%
group_6283	13%	6%
group_6291	25%	24%
group_6292	69%	79%
group_630	28%	15%
group_6304	25%	9%
group_631	22%	24%
group_6317	16%	15%
group_6322	25%	3%
group_6328	9%	18%
group_633	22%	24%
group_6331	3%	21%
group_6353	6%	18%
group_6355	6%	18%
group_6363	9%	15%
group_6367	9%	12%
group_6368	13%	21%
group_6372	3%	12%
group_6373	3%	12%
group_6374	3%	12%
group_6375	3%	12%
group_6377	3%	12%
group_6378	3%	12%
group_6379	3%	12%
group_6388	3%	18%
group_64	25%	32%
group_6408	41%	15%
group_6420	13%	9%
group_6435	6%	9%
group_6454	6%	3%
group_6460	3%	6%
group_6462	3%	6%
group_6464	9%	18%
group_6465	9%	15%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_6485	9%	0%
group_6488	9%	0%
group_6491	13%	6%
group_6492	13%	6%
group_6509	3%	9%
group_6522	0%	9%
group_6523	9%	6%
group_657	19%	9%
group_66	28%	24%
group_665	84%	94%
group_67	19%	26%
group_6728	28%	0%
group_678	34%	21%
group_6843	9%	15%
group_686	9%	21%
group_687	19%	50%
group_688	25%	3%
group_689	13%	26%
group_696	6%	15%
group_6964	6%	29%
group_6965	6%	29%
group_6966	6%	26%
group_6967	6%	29%
group_6968	6%	29%
group_6980	13%	15%
group_6983	25%	44%
group_6984	25%	44%
group_6985	25%	44%
group_6986	25%	44%
group_6997	69%	0%
group_7024	19%	44%
group_7031	63%	12%
group_7055	28%	41%
group_7089	19%	15%
group_7095	28%	21%
group_7096	28%	21%
group_7097	28%	21%
group_7098	28%	21%
group_7099	28%	21%
group_7100	6%	15%
group_7106	44%	44%
group_7133	6%	18%
group_7135	13%	18%
group_7137	16%	12%
group_7138	25%	0%
group_7174	13%	15%
group_7175	13%	15%
group_7179	81%	79%
group_7185	84%	56%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_720	22%	38%
group_7222	69%	79%
group_7227	9%	15%
group_7229	9%	18%
group_7230	9%	18%
group_7233	9%	18%
group_7234	9%	18%
group_7235	9%	18%
group_7236	9%	18%
group_7237	9%	18%
group_7238	9%	18%
group_7239	9%	21%
group_7243	9%	24%
group_725	88%	85%
group_7254	25%	18%
group_726	22%	15%
group_7262	22%	6%
group_7263	31%	0%
group_7264	31%	0%
group_7266	31%	44%
group_7267	31%	44%
group_7268	31%	44%
group_7269	31%	44%
group_7270	28%	35%
group_7272	13%	15%
group_7276	31%	26%
group_7282	66%	29%
group_7283	66%	79%
group_7284	66%	79%
group_7287	66%	35%
group_7288	66%	35%
group_7289	66%	35%
group_7291	66%	35%
group_7296	72%	44%
group_7300	47%	3%
group_7301	31%	9%
group_7306	31%	21%
group_733	19%	6%
group_7334	22%	6%
group_7341	31%	18%
group_7351	25%	0%
group_7352	25%	0%
group_7362	28%	21%
group_739	28%	12%
group_7404	19%	26%
group_7405	19%	26%
group_7406	66%	6%
group_742	53%	29%
group_7439	28%	3%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_744	19%	18%
group_7480	31%	18%
group_7482	31%	24%
group_749	56%	50%
group_750	19%	32%
group_7514	13%	41%
group_7520	72%	82%
group_7527	34%	38%
group_7542	13%	3%
group_7543	13%	3%
group_7544	13%	3%
group_755	19%	21%
group_756	3%	12%
group_7562	66%	100%
group_7563	25%	6%
group_7565	41%	0%
group_7570	9%	18%
group_7579	59%	32%
group_7580	31%	35%
group_7581	94%	85%
group_7608	19%	3%
group_7613	31%	35%
group_7627	25%	6%
group_7628	25%	6%
group_7630	34%	6%
group_7640	19%	44%
group_7648	25%	0%
group_7649	19%	6%
group_765	16%	15%
group_7655	25%	21%
group_7657	13%	32%
group_766	44%	35%
group_768	41%	0%
group_7687	6%	18%
group_77	38%	15%
group_7731	44%	9%
group_774	28%	3%
group_776	13%	12%
group_777	9%	15%
group_7789	25%	44%
group_7795	16%	18%
group_78	22%	26%
group_7801	22%	9%
group_7805	19%	9%
group_7806	25%	18%
group_7818	22%	15%
group_7826	28%	6%
group_7829	22%	6%
group_7837	59%	26%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_7839	59%	26%
group_784	28%	3%
group_7850	25%	0%
group_7854	22%	3%
group_7858	25%	0%
group_7864	34%	65%
group_7865	25%	26%
group_7885	31%	3%
group_7886	38%	6%
group_7892	38%	6%
group_7896	16%	15%
group_7909	31%	88%
group_792	25%	3%
group_793	13%	9%
group_7934	31%	24%
group_7936	31%	38%
group_7937	31%	65%
group_7953	31%	6%
group_7957	75%	82%
group_7958	22%	3%
group_7964	38%	6%
group_7968	47%	82%
group_7972	19%	6%
group_7978	28%	18%
group_7996	41%	91%
group_7997	41%	91%
group_7999	22%	6%
group_8012	22%	0%
group_8013	22%	3%
group_8015	25%	94%
group_8017	19%	6%
group_8020	59%	15%
group_8032	22%	12%
group_8037	25%	3%
group_8049	16%	12%
group_806	6%	32%
group_8076	16%	9%
group_8078	13%	15%
group_8108	47%	71%
group_811	16%	6%
group_812	16%	12%
group_813	9%	12%
group_8157	16%	9%
group_816	19%	15%
group_8162	41%	3%
group_8163	41%	3%
group_8172	38%	3%
group_8224	25%	32%
group_8248	13%	12%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_8257	13%	12%
group_8259	13%	12%
group_8260	13%	12%
group_8261	13%	12%
group_8274	16%	44%
group_8308	16%	50%
group_8322	0%	24%
group_833	31%	21%
group_8333	13%	15%
group_835	19%	12%
group_836	44%	6%
group_8361	6%	18%
group_8362	6%	18%
group_8364	9%	24%
group_8365	9%	24%
group_8369	13%	41%
group_8372	0%	6%
group_8376	0%	21%
group_839	28%	12%
group_8396	34%	88%
group_8407	0%	15%
group_8408	0%	21%
group_8409	0%	21%
group_8410	0%	21%
group_8416	28%	15%
group_8422	31%	50%
group_843	28%	9%
group_844	38%	26%
group_845	25%	9%
group_8465	6%	21%
group_8466	6%	18%
group_8467	6%	18%
group_8468	6%	18%
group_847	38%	9%
group_8473	6%	18%
group_8478	3%	24%
group_8497	13%	38%
group_8498	9%	18%
group_8528	16%	35%
group_8534	3%	26%
group_8535	0%	21%
group_8536	0%	21%
group_8537	0%	21%
group_8538	0%	21%
group_8539	0%	21%
group_8540	0%	21%
group_8551	3%	15%
group_8598	22%	0%
group_860	50%	21%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_863	6%	3%
group_8682	28%	56%
group_8688	3%	3%
group_8693	9%	9%
group_8770	31%	71%
group_882	19%	32%
group_8824	0%	9%
group_883	28%	68%
group_885	13%	24%
group_887	22%	91%
group_888	25%	29%
group_8906	13%	9%
group_8907	13%	9%
group_891	13%	21%
group_892	19%	18%
group_894	9%	15%
group_895	6%	18%
group_898	22%	9%
group_899	25%	38%
group_9017	13%	26%
group_9029	19%	29%
group_9063	3%	18%
group_912	25%	26%
group_916	3%	15%
group_9178	6%	18%
group_920	25%	9%
group_921	25%	3%
group_9229	34%	35%
group_9230	3%	26%
group_9231	0%	26%
group_9232	3%	26%
group_9233	3%	21%
group_924	25%	38%
group_926	34%	18%
group_9276	28%	0%
group_9311	22%	29%
group_937	41%	21%
group_94	16%	18%
group_940	16%	9%
group_9400	28%	38%
group_9401	28%	38%
group_9402	28%	38%
group_9410	28%	26%
group_942	44%	21%
group_9446	28%	26%
group_9447	28%	26%
group_9448	28%	26%
group_945	91%	85%
group_9452	28%	38%

Gene	Cluster 1 (%)	Cluster 2 (%)
group_9464	28%	26%
group_947	53%	50%
group_948	94%	79%
group_9484	28%	0%
group_9485	31%	32%
group_9486	31%	32%
group_9488	28%	26%
group_9503	31%	0%
group_9523	0%	47%
group_9524	0%	47%
group_953	16%	38%
group_954	50%	6%
group_9561	9%	18%
group_9563	9%	18%
group_9564	9%	18%
group_9565	9%	18%
group_9568	9%	18%
group_957	22%	26%
group_9570	9%	18%
group_9571	9%	18%
group_9573	9%	18%
group_9574	9%	18%
group_961	28%	6%
group_963	22%	3%
group_965	78%	35%
group_9651	6%	6%
group_967	6%	15%
group_9670	9%	12%
group_969	72%	21%
group_970	13%	21%
group_9704	19%	9%
group_974	28%	29%
group_9745	3%	6%
group_9790	31%	21%
group_9805	6%	24%
group_9807	19%	12%
group_9809	28%	6%
group_984	28%	38%
group_9847	13%	9%
group_9848	13%	9%
group_9870	25%	0%
group_9873	25%	0%
group_9921	16%	21%
group_997	9%	18%
gspA	72%	71%
gspE_3	6%	18%
gspI	69%	74%
gspO	69%	74%
hcpA	41%	91%

Gene	Cluster 1 (%)	Cluster 2 (%)
hha_2	19%	24%
higA_1	100%	76%
higA_2	69%	91%
higA_3	22%	18%
higA-2	78%	82%
higB_1	44%	44%
higB-1	88%	76%
higB-2_1	88%	94%
higB-2_2	78%	79%
hipA	47%	65%
hipA_1	13%	15%
hipB	72%	76%
hns_2	31%	35%
hofQ_2	6%	18%
hokA_1	28%	15%
hokA_2	9%	21%
hokD	19%	3%
hokE_2	47%	53%
hokE_3	28%	15%
hscC	41%	44%
hsdM	16%	15%
hsdM_1	3%	3%
hsdR	47%	32%
hsdR_1	3%	3%
hsdR_2	16%	6%
htrL	38%	3%
hxlB_1	31%	68%
hyfR_1	81%	94%
hyi_2	53%	82%
icd_2	72%	62%
idnD	63%	82%
idnK	63%	82%
idnO_2	31%	38%
idnR	63%	82%
idnT	63%	82%
imm_1	28%	59%
imm_2	34%	74%
imm_3	25%	53%
insA-1	69%	79%
insAB-1_1	22%	3%
insC-1	28%	12%
insC-1_2	31%	21%
insCD-1	28%	15%
insE-1	9%	15%
insEF-1	16%	15%
insG	63%	100%
insG_2	16%	35%
insJ	66%	88%
insL-3	6%	6%

Gene	Cluster 1 (%)	Cluster 2 (%)
insN-1	28%	26%
insO-1_1	16%	12%
insO-2	3%	21%
intA_2	9%	18%
intB_1	25%	24%
intB_4	34%	3%
intB_5	13%	0%
intD_2	19%	44%
intE	56%	24%
intE_1	28%	6%
intE_2	22%	32%
intQ	25%	6%
intR	22%	3%
intS	19%	9%
intS_2	6%	18%
intZ	34%	53%
intZ_1	22%	24%
intZ_2	19%	6%
kdsB_2	25%	94%
kdsD_2	81%	91%
kdsD_3	25%	94%
kilR	56%	47%
kpsM	25%	88%
kptA_1	22%	18%
lacA_1	19%	12%
lacY_2	34%	32%
lamB_1	41%	41%
ldhA_2	31%	6%
ldrB	13%	24%
legI	22%	12%
lexA_2	22%	21%
lexA_3	25%	0%
lgrD	38%	6%
lgrE	38%	6%
lolA_2	6%	18%
lomR_2	9%	18%
lpd_3	53%	82%
lsrB	59%	26%
lsrR_1	6%	29%
macA	47%	59%
macB	34%	62%
malK_1	22%	12%
malK_2	0%	56%
malX_2	81%	91%
malX_3	19%	6%
malY_2	69%	79%
manX_4	31%	6%
manX_5	31%	6%
manZ_3	31%	6%

Gene	Cluster 1 (%)	Cluster 2 (%)
marB	78%	97%
mbeC	22%	26%
mbtB	75%	74%
mcbR	88%	65%
mchI	3%	18%
mcrB	3%	6%
mcrC	3%	6%
mdtD	63%	56%
mec	78%	41%
mfd	69%	82%
mhpA	19%	0%
mhpB	19%	0%
mhpC	19%	0%
mhpD	19%	0%
mhpE	19%	0%
mhpF	19%	0%
mhpR	19%	0%
mleN	53%	82%
mngA_1	0%	15%
mngB	0%	15%
mngR	63%	53%
mngR_1	0%	15%
mobA_2	13%	6%
mokB	69%	29%
motA_2	19%	32%
mprA_2	59%	88%
mprA_3	22%	3%
mrr	16%	15%
msbA_4	28%	18%
murR_3	19%	6%
nadB_1	38%	44%
nadR_1	34%	12%
nanE_2	19%	6%
nanR_2	13%	26%
narZ	81%	91%
neuC	16%	12%
nmpC	63%	50%
nmpC_1	31%	53%
nohA	22%	3%
nohB	66%	24%
npr_2	63%	53%
ntdC	13%	6%
nupX	6%	24%
ogl	31%	38%
ogrK_1	19%	44%
ogrK_2	31%	44%
ompC_2	6%	18%
ompF	78%	79%
ompF_2	19%	6%

Gene	Cluster 1 (%)	Cluster 2 (%)
ompL	66%	35%
ompW	75%	56%
paaF_1	63%	12%
paaH	38%	6%
pagN	19%	9%
papB_1	16%	9%
papB_2	25%	18%
parA	9%	0%
parM	22%	12%
parM_1	22%	18%
parM_2	19%	26%
pduC	34%	88%
pduD	34%	88%
pduE	34%	88%
pduL	34%	88%
pelX	16%	29%
pgaA	44%	94%
pgaB	41%	65%
pgaC	22%	47%
pgaD	44%	94%
pgtC	25%	38%
phoE_1	69%	79%
phoH_1	47%	32%
php_2	38%	3%
pinE	3%	12%
pinE_1	16%	29%
pinQ	3%	6%
pinR	13%	26%
pksR	25%	24%
pld	9%	15%
pls	6%	18%
pnuC_2	34%	18%
potE	47%	71%
pptA	44%	0%
prlF	69%	88%
proA_1	13%	26%
prpC	72%	85%
prpR	59%	74%
prsE	28%	29%
prsF	13%	38%
prtR	34%	24%
ptlE	3%	6%
ptrA	56%	74%
ptsI_2	53%	68%
pys2	44%	94%
rayT	47%	68%
rbbA	72%	76%
rbsB_1	53%	88%
rbsC_3	59%	26%

Gene	Cluster 1 (%)	Cluster 2 (%)
rbsK_3	44%	47%
rbsK_4	56%	6%
rcnA	84%	88%
rcsB	25%	24%
recE	9%	12%
recF_2	13%	0%
recQ_1	41%	32%
relE_1	3%	6%
relE2	19%	21%
rem	16%	6%
renD	16%	18%
rep_1	16%	21%
rep_2	16%	15%
repB	72%	26%
rfaE_2	16%	29%
rfaH_2	9%	18%
rfaL	44%	88%
rfbA	56%	44%
rfbB	59%	50%
rfbD	59%	50%
rhsA	28%	0%
rhsB	31%	0%
rop	22%	29%
rpe_2	31%	68%
rpiB_2	31%	71%
rpiB_3	31%	35%
rrrD	25%	44%
rrrD_1	38%	9%
rrrD_2	19%	9%
rrrQ	25%	41%
rrrQ_1	19%	6%
rsmA_2	50%	24%
rsmI_2	9%	15%
rspA_2	91%	79%
rspB_2	34%	65%
rsxC_2	31%	82%
rusA	63%	50%
rusA_1	28%	15%
rusA_2	16%	18%
rusA_3	13%	9%
rutG_2	25%	6%
rzpD	41%	15%
rzpD_1	28%	6%
rzpQ	25%	0%
sacA	34%	32%
safA	56%	41%
scpA	84%	85%
scrB	38%	41%
scrY	53%	65%

Gene	Cluster 1 (%)	Cluster 2 (%)
selD	59%	56%
sfmH	94%	85%
sfmM2	13%	18%
sgcA	44%	29%
sgcA_1	41%	3%
sgcB	41%	3%
sgcB_2	31%	71%
sgcC	41%	3%
sgcE	41%	3%
sgcQ	41%	3%
sgcR	41%	3%
sgcX	41%	3%
shiA_2	34%	65%
shlB	31%	35%
smc	53%	12%
sohB_2	56%	26%
soj	34%	9%
speF	47%	65%
spo0C	38%	18%
srp54	22%	12%
ssb_2	22%	15%
ssb_4	25%	24%
stfE	13%	18%
stfR	19%	26%
stfR_2	19%	32%
sucA_1	97%	79%
symE	59%	88%
tam	94%	88%
tcpE	6%	18%
tfaD	53%	18%
tfaD_2	19%	15%
tfaE_2	19%	9%
tfaQ	38%	15%
tktB	91%	88%
tktB_2	6%	29%
tnpA_1	16%	12%
tnpA_2	16%	12%
tonB_2	59%	18%
torI	22%	6%
torR_5	84%	85%
torZ	84%	91%
traA	38%	24%
traC	9%	18%
traC_1	13%	18%
traD	44%	24%
traG	3%	6%
traI	41%	24%
traM	13%	18%
traY	16%	12%

Gene	Cluster 1 (%)	Cluster 2 (%)
treB_1	41%	41%
trpE_2	56%	59%
tsx_2	59%	71%
tufA	59%	44%
ugd	50%	62%
ulaA_2	47%	29%
ulaB_2	44%	29%
umuC_2	16%	24%
umuD_2	25%	26%
ushA_3	91%	85%
vapB	19%	9%
vapC	22%	18%
vgrG1	47%	0%
vgrG1_1	31%	24%
vgrG1_2	28%	32%
vgrG1_3	6%	15%
vgrG1_4	6%	32%
virB	34%	9%
virB1	3%	6%
virB4	3%	6%
virB9	3%	6%
waaI	34%	50%
waaJ_1	56%	97%
waaJ_2	56%	71%
waaY	59%	97%
wcaA_2	19%	29%
wzc	78%	76%
wzzB	72%	53%
xerC_1	19%	9%
xerC_3	22%	32%
xerD_2	69%	26%
xerD_3	25%	0%
xisE	78%	47%
xylG_1	6%	24%
yaaU_2	34%	65%
yadC	81%	59%
yadD	72%	65%
yadD_1	19%	15%
yadD_2	19%	24%
yadM	88%	94%
yaeF	72%	94%
yafN	47%	53%
yafO	44%	53%
yafP	91%	79%
yafT	31%	65%
yafX	19%	21%
yafX_1	41%	6%
yafX_2	22%	3%
yagA	38%	6%

Gene	Cluster 1 (%)	Cluster 2 (%)
yagG	28%	38%
yagI	69%	79%
yagM	31%	41%
yahA_2	6%	32%
yahF_1	34%	56%
yahJ	41%	68%
yahK	59%	74%
yaiF_2	44%	21%
yaiL	88%	82%
yaiO	94%	85%
yaiO_2	28%	3%
yaiP	69%	82%
yaiU_2	69%	91%
yaiW	88%	91%
yaiX	88%	76%
yajR	69%	65%
ybcC	53%	26%
ybcK	16%	18%
ybcN	25%	53%
ybcO	9%	56%
ybcQ	9%	50%
ybcW	19%	3%
ybeF	59%	59%
ybfL	28%	32%
ybfL_2	16%	26%
ybfQ	31%	29%
ybgD_1	25%	47%
ybgO_1	75%	44%
ybiP_2	22%	15%
ycaO	66%	68%
ycbF	94%	59%
ycdT	25%	50%
ycf3	50%	47%
ycfK	19%	15%
ycgV	66%	59%
ycgV_2	28%	21%
yciF	72%	56%
yciG	31%	21%
yciT	88%	85%
ycjM	81%	94%
ycjO	84%	97%
ycjP	84%	100%
ycjQ	84%	100%
ycjR	84%	100%
ycjS	84%	97%
ydaT	22%	24%
ydaV	19%	21%
ydaV_1	13%	9%
ydaV_2	38%	3%

Gene	Cluster 1 (%)	Cluster 2 (%)
ydbC	44%	71%
ydcM	50%	76%
ydcM_2	69%	9%
yddA	72%	29%
yddB	72%	29%
ydeI	19%	9%
ydeR	41%	44%
ydeS	88%	85%
ydfA	44%	6%
ydfB	44%	6%
ydfD	34%	6%
ydfE	28%	3%
ydfJ_2	63%	12%
ydfK	19%	24%
ydfN	38%	12%
ydfR	6%	24%
ydfR_1	19%	6%
ydfT	13%	18%
ydfT_2	28%	6%
ydfU	6%	18%
ydfU_1	34%	6%
ydfU_2	25%	6%
ydfX	16%	6%
ydjH_3	13%	26%
yeaM_2	78%	94%
yedK_2	16%	9%
yeeJ_2	34%	47%
yeeJ_3	81%	88%
yeeJ_4	78%	76%
yeeO_1	38%	6%
yeeP	69%	41%
yeeP_1	25%	21%
yeeP_3	9%	12%
yeeR	16%	15%
yeeS	63%	9%
yeeS_1	22%	44%
yeeS_2	22%	9%
yeeT_1	13%	26%
yeeT_2	31%	26%
yeeT_3	69%	9%
yeeW	63%	9%
yeeW_1	28%	24%
yegD	91%	82%
yegE	78%	82%
yehA	75%	68%
yehB	88%	79%
yehK	84%	74%
yehL	84%	91%
yeiL	91%	68%

Gene	Cluster 1 (%)	Cluster 2 (%)
yejO_2	69%	71%
yfaL_2	6%	12%
yfaQ	97%	85%
yfaV_2	13%	26%
yfaW	94%	82%
yfaX	88%	88%
yfbK	94%	62%
yfbS_2	69%	88%
yfcO	69%	35%
yfcS_2	13%	38%
yfdM	25%	15%
yfdM_1	41%	9%
yfdN	25%	24%
yfdO	22%	24%
yfdO_1	25%	26%
yfdO_2	22%	3%
yfdO_4	25%	9%
yfdP	22%	21%
yfdQ	28%	18%
yfdR	34%	18%
yfdS	31%	24%
yfdT	19%	9%
yfeA	41%	82%
yfgF	75%	65%
yfjI	9%	15%
yfjJ	16%	3%
yfjP_1	31%	6%
yfjQ_1	25%	15%
yfjQ_2	22%	35%
yfjQ_3	19%	9%
yfjQ_5	22%	3%
yfjR	16%	3%
yfjT	28%	38%
yfjX	25%	18%
yfjX_1	22%	24%
yfjX_2	16%	21%
ygcE	75%	88%
ygcG	47%	6%
ygcG_1	31%	29%
ygcG_2	31%	65%
ygcG_3	31%	74%
ygeV_2	25%	41%
ygfI	81%	94%
yggF	78%	79%
yggR_2	9%	18%
yghJ	78%	74%
ygiL	88%	59%
ygiS	38%	79%
yhaV	69%	88%

Gene	Cluster 1 (%)	Cluster 2 (%)
yhbO	91%	82%
yhcR	50%	21%
yhdJ_2	19%	24%
yhfK	50%	59%
yhfX	72%	100%
yhgA	56%	44%
yhgA_2	22%	24%
yhgE	88%	79%
yhjB_2	38%	6%
yhjH_2	9%	12%
yhjV	47%	88%
yiaD_1	41%	76%
yiaG_1	88%	94%
yiaN_3	41%	35%
yiaO_3	63%	94%
yibD_1	31%	29%
yidJ_2	31%	3%
yieH	78%	97%
yigE	66%	100%
yihF	81%	71%
yihL	9%	24%
yihN	53%	82%
yihO	66%	35%
yihP	66%	35%
yihQ	66%	35%
yihR	66%	35%
yihS	66%	32%
yihT	34%	65%
yihU	34%	65%
yihV	34%	65%
yihV_1	41%	41%
yihW	34%	65%
yijO_2	34%	88%
yjcE_2	66%	100%
yjcS	75%	91%
yjdJ	94%	82%
yjeN	56%	18%
yjgB	63%	82%
yjhF	41%	3%
yjhG	41%	3%
yjhH	38%	3%
yjhI	41%	3%
yjhP_1	41%	3%
yjhQ	41%	3%
yjhR	41%	3%
yjhU	31%	71%
yjiA_2	22%	6%
yjiR	31%	24%
yjiY	72%	97%

Gene	Cluster 1 (%)	Cluster 2 (%)
yjjJ	31%	24%
yjjL_1	31%	0%
yjjM	66%	97%
yjjN_2	66%	100%
yjjQ	81%	85%
ykfA	16%	3%
ykfF	25%	24%
ykfF_1	28%	18%
ykfF_2	28%	38%
ykfG	13%	3%
ykfI	13%	3%
ylbG	6%	3%
ylcG	59%	6%
yliE_2	13%	29%
ylpA	56%	24%
ymfK	13%	24%
ymfK_1	25%	6%
ymfK_2	16%	12%
ymfL	31%	29%
ymfM	31%	9%
ymfN	28%	15%
ymfR	31%	15%
ymfT	16%	24%
ynaA	47%	15%
ynaA_1	34%	0%
ynbA	44%	71%
ynbB	44%	65%
ynbC	28%	38%
ynbD	47%	71%
yncG_1	31%	94%
yncG_2	31%	94%
yneF	50%	82%
yneH	66%	76%
ynfB	72%	94%
ynfE	56%	65%
ynfF	72%	76%
ynfG	59%	74%
ynfH	69%	76%
yoeA	50%	3%
yojI	84%	59%
ypdI	63%	91%
ypjA	94%	71%
ypjA_2	19%	12%
yqaB	59%	68%
yqcE	75%	68%
yqiG	88%	59%
yqiG_2	16%	44%
yqiH	88%	59%
yqiK	63%	62%

Gene	Cluster 1 (%)	Cluster 2 (%)
yraI	34%	35%
yrhB	53%	68%
ytfR_2	59%	26%
yvqK	31%	88%
znuA_1	59%	65%
znuA_2	75%	74%
znuB_2	69%	68%

C. Virulence factor *eae* gene association to isolates in pan genome comparison study

Summary of strains studied in comparative genomics studies (Chapter 3). Isolates depicted in red text had the *eae* virulence gene associated with them.

Strain ID	Cluster	Source	MLST, German	MLST, French	Antimicrobial Resistance	Plasmid	Virulence factors	Serotype
E4931	0	Water	3307	unknown ST	Nil	IncX1, Col156	*celb, gad, vat*	O170:H5
E4942	0	Water	3307	unknown ST	Beta-lactam	IncX1, Col156	*celb, gad, vat*	O170:H5
E5795	0	Water	3307	unknown ST	Nil	IncFIB(AP001918), IncFIC(FII)	*gad, vat*	O170:H5
E7591	0	Water	3307	unknown ST	Nil	IncX1, Col156	*celb, gad, vat*	O170:H5
E7727	0	Water	3307	unknown ST	Nil	IncX1, Col156	*celb, gad, gad, vat*	O170:H5
E9644	0	Water	1873	unknown ST	Nil	IncX1, IncFIC(FII)	*astA, gad*	:H4
B0433	0	Native Vertebrate	1873	unknown ST	Nil	IncFIB(pLF82, IncFIB)	*astA, cba, gad, iroN, iss, vat*	O62:H5
P70	0	Native Vertebrate	3307	unknown ST	Nil	No replicons found	*gad, gad, iroN, mchB, mchC, mchF, mcmA, vet*	O170:H4
W3-33	0	Native Vertebrate	355	127	Nil	IncFIB(AP001918), IncFIC(FII)	*cba, cma, gad, gad, iroN, iss, iss, mchF, tsh*	O2:H5
E2059	1	Water	95	1	Nil	IncFII(29), Col156	*gad, ireA, iss, senB, vat*	:H7
E3317	1	Water	28	300	Nil	IncFIB(AP001918), IncFII(pSE11)	*cif, eae, espA, espJ, gad, gad, nleA, nleB, tir, vat*	O177:H6

Strain ID	Cluster	Source	MLST, German	MLST, French	Antimicrobial Resistance	Plasmid	Virulence factors	Serotype
E3676	1	Water	6950	unknown ST	Nil	IncFIB(AP001918), IncFII	*gad, iss*	O173:H5
E4665	1	Water	6948	unknown ST	Nil	IncFIB(AP001918), Rep	*astA, gad, iss, pic, vat*	O104:H16
E5008	1	Water	2474	unknown ST	Nil	IncFIB(AP001918), IncFIC(FII)	*astA, gad*	O8:H10
E5456	1	Water	3290	unknown ST	Nil	No replicons found	*gad, gad*	O79:H1
E5598	1	Water	1899	36	Nil	IncFIB(AP001918), IncFII	*gad, iroN, iss, iss, mchF, vat*	O4:H40
E5623	1	Water	589	585	Nil	IncX1, Col156	*cif, eae, espA, espC, espF, espJ, gad, tir, vat*	O51:H49
E7087	1	Water	95	1	Nil	IncFII(29), Col156	*cnf1, gad, iroN, iss, senB, sfaS, vat*	O18ac:H7
E7242	1	Water	681	unknown ST	Nil	IncFIB(AP001918), IncFII	*gad, iroN, iss, iss, pic, vat*	O8:H10
E7615	1	Water	95	1	Aminoglycoside, beta-lactam, sulphonamide	IncB/O/K/Z , IncFIB(AP001918)	*gad, iroN, iss, iss, mchF, vat*	O50/O2:H7
E8621	1	Water	28	572	Nil	p0111, IncFIB (AP001918)	*cif, eae, espA, espJ, gad, gad, nleA, nleB, tir, vat*	:H6
E9303	1	Water	491	unknown ST	Nil	No replicons found	*gad, gad, iss, pic, vat*	O54:H45
E9319	1	Water	3672	315	Nil	No replicons found	*gad, iss*	O140:H14
E9472	1	Water	681	304-like	Nil	Incl1, IncFII(pHN7A8)	*astA, astA, gad*	O8:H10

Strain ID	Cluster	Source	MLST, German	MLST, French	Antimicrobial Resistance	Plasmid	Virulence factors	Serotype
29-2-Si4	1	Human	589	585-like	Aminoglycoside, beta-lactam	IncI1, ColpVC	*cif, cma, eae, espA, espC, espF, espJ, gad, gad, iroN, iss, tir, vat*	:H49
58-2-AC1	1	Human	28	unknown ST	Nil	IncY, ColpVC	*astA, cif, eae, espA, espJ, gad, nleA, nleB, tir, vat*	O96:H7
6-1-TC16	1	Human	95	1	Nil	IncFII(29), Col156	*cnf1, gad, iss, senB*	:H7
70-5-R4	1	Human	95	1	Nil	IncB/O/K/Z , IncFII(29)	*ireA, iss, senB, vat*	O1:H7
H001	1	Human	681	unknown ST	Nil	ColRNAI	*gad, pic, vat*	O8:H10
H250	1	Human	12	36	Aminoglycoside, beta-lactam, sulphonamide	Col156, IncfIB(AP001918)	*cnf1, gad, gad, ireA, iroN, iss, mchB, mchC, mchF, mcmA, senB, vat*	O4:H5
H437	1	Human	95	1	Aminoglycoside, beta-lactam, sulphonamide	IncB/O/K/Z , IncFII(29)	*gad, ireA, iss, senB, vat*	O50/O2:H7
H461	1	Human	95	1	Beta-lactam, Tetracycline	IncFII, IncFIB(AP001918)	*celb, gad, ireA, iroN, iss, iss, iss, mchF, sfaS, vat*	O18:H7
H56	1	Human	28	572	Nil	No replicons found	*cif, eae, espA, gad, gad, nleA, nleB, tir, vat*	:H6
H578	1	Human	1257	122-like	Nil	IncFIB(AP001918), IncFII(pSE11)	*gad, gad, iss, pic, vat*	O8:H10
B1716	1	Native Vertebrate	136	304-like	Nil	IncI2, IncFIB(AP001918)	*gad, gad, iroN, iss, iss, pic, vat*	O8:H10
B525	1	Native Vertebrate	6165	1	Nil	IncFIB (AP001918)	*cdtB, gad, iss, vat*	:H5

Strain ID	Cluster	Source	MLST, German	MLST, French	Antimicrobial Resistance	Plasmid	Virulence factors	Serotype
BS160	1	Native Vertebrate	unknown ST	315	Nil	IncFIB(AP001918), IncFII	*gad, iroN, iss, iss, vat*	O140:H14
M0651	1	Native Vertebrate	491	375-like	Nil	Col156, IncfII(pCoo)	*celb, gad, gad, iss, pic, vat*	O54:H45
P40	1	Native Vertebrate	3306	unknown ST	Nil	IncFIB(AP001918)	*cba, cma, gad, iss, vat*	:H45
P7	1	Native Vertebrate	1800	unknown ST	Nil	Col156, IncfIB(AP001918)	*astA, astA, cba, cdtB, celb, cma, gad, gad, iss*	:H5
TA309	1	Native Vertebrate	681	unknown ST	Nil	ColRNAI	*gad, iss, pic, vat*	O8:H10
E2038	2	Water	372	490	Nil	No replicons found	*cnf1, gad, gad, iroN, iss, mchB, mchC, mchF, mcmA, vat*	O21:H14
E2549	2	Water	1858	unknown ST	Nil	No replicons found	*gad, iroN, mchB, mchC, mchF, mcmA, pic, vat*	O6:H5
E4712	2	Water	3304	unknown ST	Nil	IncY, IncFIB(AP001918)	*gad, gad, gad, pic, vat*	:H7
E7253	2	Water	3646	unknown ST	Nil	No replicons found	*gad, gad, pic, vat*	O16:H14
E8766	2	Water	1619	331	Nil	IncX1, IncFIB(AP001918)	*cdtB, gad, iss, iss, vat*	O50/O2:H5
E2062	2	Water	3291	unknown ST	Nil	No replicons found	*gad, vat*	O25:H5
E4259	2	Water	636	700	Nil	No replicons found	*gad, gad, vat*	O83:H7
E4453	2	Water	135	unknown ST	Nil	IncFIB(AP001918), IncX1	*gad*	O83:H1
E4963	2	Water	6947	unknown ST	Nil	IncFIB(AP001918), IncX1	*gad, gad, vat*	O16:H7

Strain ID	Cluster	Source	MLST, German	MLST, French	Antimicrobial Resistance	Plasmid	Virulence factors	Serotype
E6649	2	Water	1386	unknown ST	Nil	IncFII, p0111	*astA, astA, cba, cma, gad, pic, vat*	O13:H4
E7603	2	Water	569	732	Nil	No replicons found	*gad, gad, iss, iss, vat*	O134:H31
E8279	2	Water	6949	unknown ST	Nil	no replicons found	*gad, vat*	O150:H5
E9345	2	Water	95	unknown ST	Nil	p0111, IncFIB (AP001918)	*cba, cdtB, cma, gad, iroN, iss, sfaS, vat*	O120:H5
E9693	2	Water	1925	unknown ST	Nil	ColRNAI	*gad, ireA, mchB, mchC, mchF, mcmA, vat*	O170:H14
E9866	2	Water	2800	unknown ST	Nil	Incl1	*gad, vat*	O46:H7
DMG-2015	2	Human	2800	unknown ST	Nil	Incl1, Incl1	*astA, gad, vat*	O46:H7
H588	2	Human	126	unknown ST	Nil	No replicons found	*gad, gad, gad, vat*	:H5
B351	2	Native Vertebrate	127	unknown ST	Nil	No replicons found	*gad, gad, pic, vat*	O75:H31
M652777	2	Native Vertebrate	372	455	Nil	No replicons found	*gad, gad, iss, pic, vat*	:H31
B015	2	Native Vertebrate	126	unknown ST	Nil	Col(MG828)	*astA, gad, vat*	:15
B103	2	Native Vertebrate	1894	unknown ST	Nil	p0111, IncFIB (AP001918)	*cba, cma, gad, gad, vat*	O13:H5
B1587	2	Native Vertebrate	6949	unknown ST	Nil	Incl1	*gad, vat*	O150:H5
B620	2	Native Vertebrate	2622	unknown ST	Nil	p0111, ColRNAI	*astA, gad, gad, vat*	O83:H6

Strain ID	Cluster	Source	MLST, German	MLST, French	Antimicrobial Resistance	Plasmid	Virulence factors	Serotype
M0528	2	Native Vertebrate	1858	51	Nil	No replicons found	*gad, gad, pic, vat*	O75:H5
M0540	2	Native Vertebrate	1925	unknown ST	Nil	IncFII, ColRNAI	*cba, cma, gad, ireA, mchB, mchC, mchF, mcmA, vat*	O170:H14
M619443	2	Native Vertebrate	6949	unknown ST	Nil	No replicons found	*gad, gad, vat*	O150:H5
M660200	2	Native Vertebrate	6998	unknown ST	Nil	No replicons found	*gad, iss, vat*	:H1
M694984	2	Native Vertebrate	569	732-like	Nil	IncFII(pCoo)	*cba, cma, gad, iss, pic, vat*	O134:H31
TA025	2	Native Vertebrate	126	unknown ST	Nil	No replicons found	*gad, vat*	:H5
TA059	2	Native Vertebrate	6952	unknown ST	Nil	Col156, IncfII(pCoo)	*cba, cma, gad, ireA, mchB, mchC, mchF, mcmA, pic, vat*	O75:H7
TA206	2	Native Vertebrate	1386	unknown ST	Nil	IncFIB(AP001918), Col156	*cba,celb, cma, gad, gad, gad, pic, sfaS, vat*	O13/O135: H4
TA209	2	Native Vertebrate	5430	unknown ST	Nil	Col156, ColRNAI	*celb, gad, gad, iroN, iss, mchB, mchC, mchF, mcmA, vat*	O75:H7
TAS-4617461	2	Native Vertebrate	135	unknown ST	Nil	IncFIB(AP001918), IncFII	*gad, iroN, iss, vat*	O2:H1
W2-51	2	Native Vertebrate	135	unknown ST	Sulphonamide	Col8282, ColpVC	*astA, astA, astA, astA, cma, gad, iroN, iss, mchF, tsh*	O2:H1

D. Top three genes found to explain the most variation observed in APW and FPW

D.1 Nucleotide sequence of the top three genes that affected variation in isolates days to recovery from VBNC state in APW

> ariR
ATGCTTGAAGATACTACTATTCATAATGCAATATCTGATAAAGCGTTATCAAGTTACTTT
CGTAGTTCAGGTAATTTGTTAGAAGAAGAGTCAGCTGTGTTAGGGCAGGCTGTGACCAAT
TTAATGCTTTCAGGTGATAATGTTAATAATAAAAATATTATCTTAAGTCTGATACACTCC
CTGGAAACAACAAGTGATATTCTCAAAGCTGATGTGATTAGAAAAACACTGGAAATCGTG
TTGCGATACACAGCTGATGATATGTAA

> gatA_2
ATGCAAGGCATACAGTTTCAGGAAAATTTTATTCAACGACTTCCGGCAGGGTTAAGCGTT
GAACAGATCATCCGTCAGTTAGCGCAACCGTTGGTTACCGCCGAACTGGTGGTCCCTGAT
TTTGCCGATCATGTACTGGAACGTGAGGCGACGTACCCAACGGGACTGCCTACCGAACCA
CCGTGTGTCGCCATTCCGCACACGGACCATAAACATGTCAGGCACAATGCAATTGCCGTC
GGCATCCTGCCGGAACCGGTAGTGTTTGCCGATATGGGCGGCGACCCTGATCCCGTACCT
GTCAGGGTGATTTTTTTGCTCGCCTTAGGCGAAAGTAACAAACAATTAAATGCCCTGGGG
TGGATTATGGAGATGATTCAGGATACGCCCTTTATGTGCGCCCTGCTTACGATGGAAACA
ACAGAAATACACACAGCAATCTTAAACAAAATGAAAGAACGAGGTGAAATATGA

> group_40
ATGGGTAAAGGCAGCAGTAAGGGGCATACCCCGCGCGAAGCGAAGGATAACCTGAAATCC
ACGCAGTTACTGAGTGTGATCGATGCCATCAGCGAAGGGCCGATTGAAGGTCCGGTGGAT
GGATTAAAAAGCGTGCTGCTGAACAGTACACCGGTGCTGGACAGTGAGGGGAATACCAAC
ATCTCCGGTGTCACGGTGGTGTTCCAGGCAGGTGAGCAGGAGCAGACACCGCCGGAGGGA
TTTGAATCCTCCGGCTCCGAGACGGTGCTGGGTACGGAAGTGAAATATGACACGCCGATC
ACCCGGACCATCACGTCGGCAAACATTGACCGTCTGCGCTTTACCTTCGGTGTGCAGGCA
CTGGTGGAAACCACCTCAAAGGGGGACCGGAATCCGTCGGAAGTTCGCCTGCTGGTTCAG
ATACAGCGTAATGGTGGCTGGGTGACGGAAAAAGACATCACCATTAAGGGCAAAACCACC
TCGCAGTATCTGGCCTCGGTGGTGGTGGATAACCTGCCGCCGCGCCCGTTTAATATCCGG
ATGCGCAGAATGACGCCGGACAGCACCACAGACCAGCTGCAGAACAAAACGCTCTGGTCG
TCATACACCGAAATCATCGATGTGAAACAGTGCTACCCGAACACGGCACTGGTCGGCGTG
CAGGTGGATTCGGAGCAGTTCGGCAGCCAGCAGGTGAGCCGTAATTATCATCTGCGCGGG
CGTATTCTGCAGGTGCCGTCGAATTATAACCCGCAGACGCGGCAATACAGCGGTATCTGG
GACGGAACGTTTAAACCGGCATACAGCAACAACATGGCCTGGTGTCTGTGGGATATGCTG
ACCCATCCGCGCTACGGCATGGGGAAACGTCTTGGTGCGGCGGATGTGGATAAATGGGCG
CTGTATGTCATCGGCCAGTACTGCGACCAGTCAGTGCCGGACGGCTTTGGCGGCACGGAG
CCGCGCATCACCTGTAATGCGTACCTGACCACACAGCGCAAGGCGTGGGATGTGCTCAGT
GATTTCTGCTCGGCGATGCGCTGTATGCCGGTATGGAACGGGCAGACGCTGACGTTCGTG
CAGGACCGACCGTCGGATAAGGTGTGGACCTATAACCGCAGTAATGTGGTGATGCCGGAT
GATGGCGCGCCGTTCCGCTACAGCTTCAGCGCCCTGAAAGACCGCCATAATGCCGTTGAG
GTGAACTGGATTGACCCGAACAACGGCTGGGAGACGGCGACAGAGCTTGTGGAGGATACG

CAGGCCATTGCCCGTTACGGTCGTAACGTCACGAAGATGGATGCTTTTGGCTGTACCAGC
CGGGGGCAGGCACACCGCGCCGGGCTGTGGCTGATTAAAACAGAACTGCTGGAAACGCAG
ACCGTGGACTTCAGCGTGGGCGCAGAAGGGCTTCGCCATGTGCCGGGCGATGTCATTGAA
ATCTGTGATGATGACTATGCCGGTATCAGCACCGGTGGTCGCGTGCTGGCGGTGAACAGC
CAGACCCGGACGCTGACGCTCGACCGTGAAATCACTCTGCCATCCTCCGGTACCACGCTG
ATAAGCCTGGTTGACGGAAGTGGCAATCCGGTCAGCGTGGAGGTTCAGTCCGTCACCGAC
GGCGTGAAGGTGAAAGTGAGCCGTGTTCCTGACGGCGTTGCTGAATACAGCGTGTGGGGG
CTGAAGCTGCCGACGCTGCGCCAGCGCCTGTTCCGCTGTGTGAGTATCCGTGAGAACGAT
GACGGTACGTATGCCATCACTGCCGTGCAGCATGTGCCGGAAAAAGAGGCCATCGTGGAT
AACGGGGCGTACTTTGACGGCGACCAGAGCGGAACGGTGAACGGTGTCACGCCGCCCGCG
GTGCAGCACCTGACTGCAGAAGTCACCGCAGACAGCGGGGAATACCAGGTGCTGGCGCGC
TGGGACACGCCGAAGGTGGTGAAGGGGGTGAGCTTTATGCTTCGCCTGACTGTGGCAGCG
GATGACGGCAGTGAGCGGCTGGTCAGCACGGCCAGGACGACGGAAACCACATACCGCTTC
AGGCAACTGGCGCTGGGGAACTACAGGCTGACAGTCCGGGCGGTAAATGCGTGGGGGCAG
CAGGGCGATCCGGTGTCGGTATCGTTCCGGATTGCCGCACCGGCAGCACCGTCGAGGATT
GATCTGACGCCGGGCTATTTTCAGATAACCGCCACGCCGCATCTTGCCGTTTATGACCCG
ACGGTACAGTTTGAATTCTGGTTCTCGGAAAAGCGGATTACCGATATCAGGCAGGTTGAA
ACCACAGCCCGCTATCTTGGCACGGCGCTGTACTGGATAGCTGCCAGTATCAATATTAAG
CCGGGCCATGATTATTATTTTTATATCCGCAGTGTGAATACTGTTGGCAAATCGGCATTC
GTGGAGGCTGTCGGTCAGGCGAGCGATGATGCTTCCGGCTATCTGGATTTTTTCAAAGGC
CAGATAACTGAATCCCATCTCGGCAAGGAGCTGCTGGAAAAAGTCGAGCTGACGGAGGAT
AACGCCAGCAAACTGGAGGAGTTTTCGAAAGAGTGGAAAGATGCCAACGATAAATGGAAT
GCCATGTGGGGCGTCAAAATTGAGCAGACCGAAGACGGCAGGCATTATGTCACGGGGCTT
GGCCTCAGCATGGAGGATACGGAGGAAGGTAAACTGAGCCAGTTCCTGGTTGCCGCTAAC
CGTATCGCGTTTATTGACCCGGCAAACGGGAATGAAACGCCGATGTTTGTGGCGCAGGGC
AACCAGATATTCATGAACGAAGTGTTCCTGAAGTATCTGACGGCTCCCACCATTACCAGC
GGCGGCAATCCGCCGGTATTTTCCCTGACACCGGACGGGCGGTTGACGGCGAAAAATGCC
GATATCAGCGGTAACGTGAATGCGAACTCCGGGACGCTCAATAATGTCACGATAAATGAG
AACTGTCAGATTAAGGGGAAACTGTCCGCCAACCAGATTGAAGGCGATATTGTCAAAACG
GTCAGCAAGTCTTTCCCCCGCACGAGCAGTTATGCCAGCGGCACCATTACGGTCACGATT
AGTGATGATCAGAAGTTTGACCGGCAGGTCATGATCCCGGCTCTGTTGTTTAAAGGAAGC
AGGAAGGAAAATTATGGCAGTAATAATCAACAGTCTTATGTTTATTCTGTATGCCGTTTG
CAGGTAACGAAAAACGGGACAGAAATTTTTAATCAGTCAACAACGGATGCTCCGGCGGTT
TTTTCTTCCGTTATTGATATGCCCGCGGGGCAGGGAACGTTGACGTTAAAATTTACGGTC
TCTTCTTCAATGGTCAATAACTGGACACCGACAACCAGTATCAGCGATTTGCTGGTTGTG
GTGATGAAGAAATCCACCGCAGGTATCACGATTAGCTGA

D.2 Nucleotide sequence of the top three genes that affected variation in isolates days to death rate in FPW

> yfdM

ATGAGTAATAAATATTGCCAGGCGCTGGTGGAACTGCGGAACAAACCAGCCCATGAACTG
AAGGAAGTGGGCGATCAGTGGCGCACGCCGGACAACATTTTCTGGGGAATTAACACCCTG
TTTGGCCCGTTTGTTCTGGATCTGTTCACTGACGGTGATAACGCCAAATGTGCCGCGTAT
TACACGGCGGAAGATAACGCGCTGGCGCATGACTGGTCAGAACGTCTTGCGGAGCTTAAA
GGTGTTGCCTTTGGTAATCCCCCATACAGCCGCGCCAGTCAGCATGAGGGGCAATACATC

ACCGGCATGCGTTACATCATGAAACATGCCAGTGCCATGCGTGATAAGGGCGGGCGCTAT
GTTTTCCTGATCAAAGCTGCCACCAGCGAAGTGTGGTGGCCGGAAGATGCGGACCATATT
GCTTTTATTCGCGGGCGTATTGGTTTTGAACTGCCTGCCTGGTTTATCCCGAAGGATGAG
AAGCAGGTGCCGACAGGAGCTTTCTTCGCTGGTGCTATTGCTGTTTTCGACAAGACCTGG
AAGGGACCGGCAATCAGCTACATCGGGCGCGATGAACTTGAGGCATGTGGTGAAGCCTTT
CTGGTGCAGGTTCGCCAGCAGGCGGAAAAACTGGTCAGGGAGATGGCGGCATGA

> *pnuC_2*
ATGTCTTATTCAGAAATCGCCGCGTGTCTGGCATACGCGGTTTCCGTCTGGCTGGCTGCC
AGAAACAATGTGCATACATGGTGGATCGGCATAATTGGCAGCATATTGTACGGCTGGGTT
TTTTGGTCCGTGCAACTCTATGCCGACGTTACGCTCCAGTTATTCTTCATCGTGACCAGC
ATCACTGGCTGGATCCACTGGCTGAAAGGTCAGGGTGGCGACATCTTGCCGGTGCGCCGA
ACGCAAGCCAGTCACTTTTTCCTTTTGTTGCTCTGTGCTGTCGTCGTGGCAGGCGGTTAC
GGCTTTTTACTCCACACCTTTACCAATGCCTGGGCACCCTGGCTGGATTCGTTGATTTTG
ACCTTCAGCGTTCTGGCACAATTCATGTTGATGGGAAGACGTATCGAAAACTGGTACGTC
TGGTTAGCGGTGAATACCCTGGCGGTGCCACTGTATATGACGCGCGGTTTAAACCTGACC
GCTGGCTTATATTTCCTGTTCTGGATTAACGCCTGGCATGGTTTGTATCAATGGCGCAAA
GAGTTGCAAACATCATGA

> *group_1212*
ATGAATACTGGATATTCTCCCGAACAAGGGCGGGGCTTCGTTCGTCCTGAAAAACAGAAT
CTGCAAAATTTTGCCGAAATTATTCCGGTTATTTCCGGCCTTACTGGCGGGAGTGAAACC
AATATTGTTAACGCCAGAGCGTTGCAGATGTTTGATGATAAAAAGGGAGTAAATTTAACT
TACACCCCTGACGGCAATCAGAATATGAGCATTATCTCTGAGTCAGGTTTCTACAAACTA
ATAAAAACAAAAAGCGCCCCGTTGCCGGAGCGCCTTTGTGAACAATTAACCTACTGCGCA
AAAAATGAATCTGAGCAGTGGGATTATATCAACCATGTGGAGAAGCGCCACAATTGCCGA
ATAACGGGCAAAACAAAGGCCACCCGCTACGGTGGCCCCTCGACACAAGCTACACGTTAT
CCCCAACGCATGAGCATTGCCAACAATGCCACATTTGCGGCTGGTGGGCAATGCAATCAG
TCAGGTTCAGTTCGTTGCCATACCTGCAATGAGCGCTTTTCCCTGTACTCTTTAAGGAAT
TGCTCAAGGGCAAAAGCACATGGCGCGAATCTTTCTGATTCATGCTCTATCTTTCTGCGC
CGTCTTTTCCGTGCCGGTGATAATGTTTTGGTCAATTCTTTATCGGTCATTGTGTTGTCC
TGCATAGCAATGCGCCGTAGTTACTCACACCACGGCGCTGGTGATGGTTACTCCTGCTCT
TTGGCCTTGCGGCGCTGGCGGCGTTTGATCTCGCCTTGCATTGACGCAATGATGAATTGT
GATGTGCTTTCGCCTGATTCTTTCACTGATTCCATGGCGTTTACTATTTCGTGTGGTACA
CGAGCCTTTAGAGCTTGTGATTTTGCGTTTGTTGTACCCGTTGCCATTTCTATTCCTCTG
TTGGATTGGTGGGGGACAGTATACACAGAAAAAGCACAAATACAACCCTTGACGTGGGCC
ACACCTAAAGCATATAGTGGGCCACACCTTGAGATTATCAAGGTGCATAAATGCAAAGCC
CCGCAAGTGTCGTTACCACTCGCAGGGCTTCTAACCACCAACGATAGCGAAAGTATCGAG
GTAGCTATGTTAAATCATACCACACACCCGCAAGGGCGGGACTCGCACAACCTGAATAAA
TACATCTGGCGTTTTATCGCCCTGAGCACGGCACAACCGCGCGTGATTACCATTGAGGCC
ACCAGCGAACAGGAAGCACGCCAGCAATCTCCGGCTGGCTGCGTGATGGTATTTGCTGCC
CGTATTCGTCAGGGGGTAGGCTTATGA